The effect of quasi-identifier characteristics on statistical bias introduced by k-anonymization

The effect of quasi-identifier characteristics on statistical bias introduced by k-anonymization

Author:

Olivia Angiuli

Advisors:

James Waldo

Joe Blitzstein

A thesis submitted in partial fulfillment of the requirements

for the joint degree of

Bachelor of Arts

in Statistics and Computer Science

Harvard College

Cambridge, Massachusetts

April 1, 2015

Abstract

The de-identification of publicly released datasets that contain personal information is necessary to preserve personal privacy. One such de-identification algorithm, k-anonymization, reduces the risk of the re-identification of such datasets by requiring that each combination of information-revealing traits be represented by at least k different records in the dataset. However, this requirement may skew the resulting dataset by preferentially deleting records that contain more rare information-revealing traits. This paper investigates the amount of bias and loss of utility introduced into an online education dataset by the k-anonymization process, as well as suggesting future directions that may decrease the amount of bias introduced during de-identification procedures.

Acknowledgements

I am most grateful for my thesis advisor Jim Waldo, who has helped direct me towards promising directions throughout the writing of this thesis. You have been an influence, not only in the writing of this thesis, but also in driving my broader interests regarding the importance of privacy in today's increasingly technical world, both from a technological and policy standpoint. I would also like to thank my advisor Joe Blitzstein for providing the initial idea of exploring de-identification from a statistical standpoint, and for your extremely valuable technical feedback on my thesis that has pushed me to think deeper about many of the discussed concepts. You have had a profound influence on my undergraduate career, including being the driving force behind my interest in Statistics.

Contents

Abstract i

Acknowledgements ii

Contents iii

List of Figures v

List of Tables vi

1 Introduction 1

2 Defining de-identification 4
2.1 Limits on the use of data . . . 5
2.2 Technical approaches toward data protection . . . 6
2.3 De-identification schemes . . . 7
2.3.1 k-anonymity . . . 7
2.3.2 l-diversity . . . 10
2.3.3 Differential privacy . . . 12
2.4 How is de-identification achieved? . . . 12
2.5 Determination of the "best" dataset for release . . . 14
2.5.1 Count Data . . . 15
2.5.2 Clustering . . . 16
2.5.3 Linear regression . . . 17
2.5.4 Classification . . . 18
2.6 Unspecialized datasets for release . . . 19

3 Experimental approach 21
3.1 Dataset description . . . 21
3.2 De-identification procedure for edX dataset . . . 22
3.3 edX data: de-identified versus original dataset . . . 26
3.4 One-dimensional changes: count data . . . 26
3.5 Two-dimensional changes: correlation data . . . 30

4 k-anonymization: identifying sources of bias 33
4.1 The relationship between quasi-identifier frequency and the bias of attributes . . . 33

4.1.1 Individual quasi-identifier frequencies 34
4.1.2 Quasi-identifier combination frequencies 40
4.2 Measures of data utility 42
4.3 The effect of k on statistical bias and utility 45
4.4 The effect of suppressing entire quasi-identifier columns on statistical bias 48
4.5 The effect of the correlation of QI rarity with grade on statistical bias . . 52
4.6 The effect of generalization versus suppression on statistical bias 54

5 Conclusions and future directions **61**

List of Figures

2.1 Illustration of dataset linking via shared quasi-identifiers 4

3.1 Changes in enrollment induced by de-identification 26
3.2 Histogram of attribute changes induced by de-identification 29
3.3 Change in attribute correlations induced by de-identification 31

4.1 Frequency of individual quasi-identifier values versus mean grade 35
4.2 Frequency of individual quasi-identifier values versus performance metrics 37
4.3 Frequency of individual quasi-identifier values versus activity metrics . . . 39
4.4 Frequency of quasi-identifier value combinations versus mean grade 40
4.5 Frequency of quasi-identifier value combinations versus performance metrics 41
4.6 Frequency of quasi-identifier value combinations versus activity metrics . 42
4.7 Change in course enrollments as k is varied 47
4.8 Change in grade, performance, and activity metrics as k is varied 48
4.9 Relationship between correlation of quasi-identifiers versus entropy 53
4.10 Correlation between rarity of quasi-identifier values versus mean grade . . 53
4.11 Change in metrics as forum post bin size is varied 56

List of Tables

2.1 Example of dataset containing personally identifiable information 5
2.2 Example of re-identification 7
2.3 Example of a k-anonymous dataset 8
2.4 Example of a (α, k)-anonymous dataset 10
2.5 Example of an l-diverse dataset 11
2.6 De-identification via swapping 16

3.1 HarvardX dataset description 22
3.2 Changes in performance and gender variables induced by de-identification 27
3.3 Chi-squared distance between original and anonymized attributes 27
3.4 Pairwise correlations for original dataset 30
3.5 Pairwise correlations for de-identified dataset 30

4.1 Frequency of individual quasi-identifier values versus performance metrics 36
4.2 Frequency of individual quasi-identifier values versus activity metrics . . . 38
4.3 Change in de-identified dataset as k is varied 46
4.4 Change in data utility as k is varied 47
4.5 Quasi-identifier frequency correlations with mean grade, activity, and performance metrics 50
4.6 Represented forum posts values as bin size is increased 55
4.7 Change in grade, performance, and activity metrics as forum posts bin size is changed 55
4.8 Change in data utility as forum post bin size is varied 57
4.9 Change in correlations as generalization increases 58
4.10 Change in data utility with generalization using mixed bin sizes 59
4.11 Change in correlations with generalization of mixed bin sizes 59

Chapter 1

Introduction

Although the prevalence of personal data collection in medical, educational, and other fields enables valuable data analysis, the public release of these datasets for research purposes has necessitated policies that protect the privacy of subjects whose information is contained in these datasets. Current legal standards have equated privacy to a notion of *anonymity*, allowing private data to be shared about people, as long as there exists a reasonable amount of uncertainty associated with the identity of such persons.

Medical data, for example, is protected by the Health Insurance Portability and Accountability Act (HIPAA), which requires that health records be cleared of highly-sensitive fields like names and telephone numbers and that the dataset be altered such that any given subject's risk of being re-identified from the released dataset is "statistically small" [1]. Similar laws also provide privacy guidelines for the release of educational, internet, financial, and other data.

Various de-identification procedures have been developed that define standards by which datasets must be altered such that a given subject's maximum risk of being re-identified falls below a given threshold. By modifying the original dataset, however, the de-identification process often results in a dataset that generates less accurate analyses than does the original data. This points to a necessary trade-off between privacy and the use of big data – the greater the required standard of anonymity, the greater the opportunity for a dataset to be altered in a way that lowers the accuracy of the resulting data analyses.

One scenario in which de-identification altered the properties of a dataset can be seen in the context of edX data, a massive open online course platform providing free classes to those who sign up online. After edX student data was de-identified in order to comply with Family Educational Rights and Privacy Act (FERPA) law, basic statistics about student activity and performance changed considerably, including the mean and correlation of various course performance characteristics [2]. Similar phenomena have also been observed with datasets in the advertising and pharmaceutical industries [3].

Optimizing the de-identification process in order to maximize the accuracy of analyses is challenging for multiple reasons. First, it usually cannot be known in advance what other datasets a given adversary may have, and therefore it is difficult to select which characteristics a dataset should be anonymized with respect to. Second, the *analyses* that researchers will perform – and therefore the attributes of the dataset whose accuracy should be optimized – are often difficult to predict in advance. Even if columns of interest *could* be predicted, optimizing for the representativeness of a certain attribute to the original dataset may come at the cost of decreasing the representativeness of other attributes, which creates distortions in correlations between attributes. One possible solution is to release multiple de-identified datasets, each optimizing for the accuracy of a specific attribute of the data. However, since the release of multiple datasets may, in combination, violate the given anonymity standard, this may necessitate each individual dataset to have an even stricter anonymity standard than if a single, optimal dataset were released.

Tiancheng Li and Ninghui Li highlight that publishing de-identified datasets represents a sacrifice of *societal* benefit for *individual* privacy gain. In this case, societal gain is represented by the knowledge obtained by performing analyses on these datasets, whereas privacy loss is defined as the information that an adversary could gain about a particular subject by joining the dataset with an outside dataset [4]. In developing and analyzing the efficacy of different de-identification procedures and requirements, the desired balance with regard to this spectrum must be taken into account.

This paper will explore the effect of a de-identification procedure, k-anonymization, on the statistical properties and relationships of the resulting dataset. The edX dataset is analyzed with the goal of identifying specific characteristics of the original dataset that may contribute to the amount of statistical bias introduced during the k-anonymization process. In this paper, we define statistical bias as the difference between a given attribute's mean in the original dataset and in the de-identified dataset. This specific case

study hopes to lend insights into the mechanisms present within de-identification methods that tend to bias data, and provides suggestions on future work that may minimize the amount of statistical bias that is introduced in these datasets during de-identification.

Chapter 2

Defining de-identification

Any dataset containing information about individual subjects will contain records (rows) with attributes (columns) that fall into at least one of the following categories [5] [6]:

Definition 2.1 (Identifier)**.** An attribute that contains explicitly identity-revealing information. Examples include a subject's name or social security number.

Definition 2.2 (Quasi-identifier)**.** Attributes that contain information that may partially reveal identity through the linking of these quasi-identifiers with external data that share the same quasi-identifiers, as illustrated in the diagram below [7]. Examples include gender or state of residence.

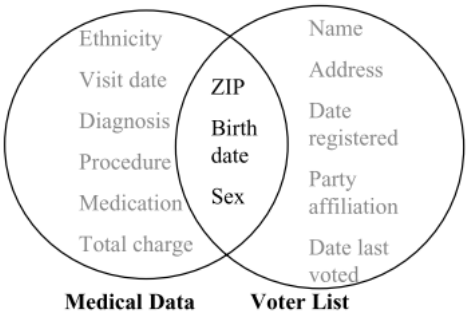


FIGURE 2.1: This diagram illustrates how quasi-identifiers can be used by an adversary in order to reveal extra information about a person by linking two datasets together.

Definition 2.3 (Sensitive attribute)**.** Attributes that can potentially cause societal or personal harm if linked with an identifier. Examples include medical diagnoses like HIV or cancer.

Definition 2.4 (Nonsensitive attribute)**.** Attributes that are not sensitive.

The below dataset illustrates examples of each of the above-described types of attributes. Social security number (SSN) is the sole identifier in this dataset, because it is the only attribute that can singly be used to identify a person from a given record. Zip code, age, and nationality are all quasi-identifiers because, if joined with an external dataset, they may be used to reveal more information about an individual corresponding to a given record. In this case, condition is a sensitive attribute because, if it were to be revealed, this would pose a privacy violation that could cause personal harm. Non-sensitive attributes are therefore social security number (SSN), zip code, age, and nationality.

	SSN	Zip Code	Age	Nationality	Condition
1	641-49-6370	13053	27	US	Heart Disease
2	929-69-9710	13053	29	Canada	Heart Disease
3	236-94-7153	13053	14	US	Viral Infection
4	926-47-9572	13053	21	Mexico	Viral Infection
5	042-48-1725	14857	42	US	Viral Infection
6	528-41-7495	14857	63	England	Heart Disease
7	925-41-0497	14857	44	China	Viral Infection
8	640-25-1430	14857	52	China	Viral Infection
9	482-55-1927	13060	33	Germany	Cancer
10	492-20-4710	13060	30	Switzerland	Cancer
11	729-46-1031	13060	39	France	Cancer
12	294-10-4528	13060	32	US	Cancer

TABLE 2.1: An example of a dataset containing personally identifiable information. SSN is an identifier. Quasi-identifiers are zip code, age, and nationality. Condition is a sensitive attribute. Nonsensitive attributes are zip code, age, and nationality [8].

Privacy protection of datasets like this may be approached from two angles: either 1. limiting the *use* of the original dataset, including restrictions either on the people who are granted access or on the methods of distribution, or 2. technical modifications that allow open access to an anonymized version of the original dataset whose risk of re-identification through linking with outside datasets is determined to be minimal. These two angles are discussed in greater detail below.

2.1 Limits on the use of data

In order to protect the privacy of individuals included in a dataset, one of the most straightforward approaches is simply to restrict access to the dataset. This may include the enforcement of strict limits regarding with whom this data can be shared and how aggregate the data findings have to be before being released. Such restrictions, however,

are often hard to enforce, especially with digital datasets that are very easily mass-distributed. Once such a privacy breach occurs, it can be nearly impossible to retract the data from people with unauthorized access. For this reason, sensitive datasets often have to be released in a controlled manner that tracks who has access to the data.

An alternative approach toward systematically preventing the abuse of sensitive data is called *dynamic anonymization* and is discussed by Xiaokui Xiao et al. Under this approach, a statistical database (StatDB) provides a user with aggregated query results against a database with personally identifiable information, such that the combination of all data provided to each given user never violates a pre-specified de-identification standard. Such a database uses one of three methods in order to answer other queries:

I. *Query restriction.* The refusal to answer queries that disclose sensitive information.

II. *Output perturbation.* The addition of noise to query result such that no sensitive exact values are compromised.

III. *Data modification.* The output of a table that has been anonymized to a pre-specified standard and that maintains the anonymity of all of the information that has been queried to date.

The strengths of such a system are its ability to manage the amount of information that each user has queried and therefore to never release information that would compromise the pre-specified anonymity level of the data. However, as a given user performs more queries, the computational cost to return new queries monotonically increases due to the need to check for adherence to de-identification standards of an increasing number of records. Furthermore, the representativeness of query results to the original dataset tends to decrease as a user queries more, since all data collected by the user to date must collectively satisfy given de-identification standards [9].

2.2 Technical approaches toward data protection

Unlike privacy procedures that focus on restricting *access* to personally identifiable or sensitive data, technical approaches toward de-identification aim to generate an "alternate" version of a given dataset with personally identifiable information such that the risk of re-identifying an individual contained in this dataset is minimal. This is accomplished through the complete removal of identifier attributes like name and SSN, along

with alterations to the dataset, as specified by each de-identification scheme, that minimize the risk of re-identification.

Re-identification most commonly occurs when data is combined with an outside source via shared quasi-identifier attributes in order to reproduce identifier attributes in the de-identified dataset. For example, consider the example medical dataset presented to the bottom left, along with the "external" Census data at the right that contains information on all three of the residents of the zip code 13053.

The two datasets, in combination, reveal that the third record in the left dataset belongs to George Leary, since his zip code and age match with that of the third record.

	Zip Code	Age	Nationality	Condition
1	13051	21	US	Heart Disease
2	13052	29	Canada	Heart Disease
3	13053	14	US	Viral Infection
4	13051	21	US	Viral Infection

Name	Zip Code	Age
John Abbott	13053	18
George Leary	13053	14
Jane Wentworth	13053	35

↓

	Name	Zip Code	Age	Nationality	Condition
3	George Leary	13053	14	US	Viral Infection

TABLE 2.2: Illustration of the re-identification that is enabled if an adversary has access to more than one dataset, one of which contains identifier attributes (such as name, in this case). The combination of datasets in this example enable George Leary to be re-identified, therefore revealing his medical condition, which was previously a sensitive attribute.

De-identification schemes attempt to minimize the chance of this re-identification. Below, some of the most common de-identification schemes are defined and explored.

2.3 De-identification schemes

2.3.1 k-anonymity

Re-identification was possible in the above example because there was only one individual in the medical dataset whose zip code and age matched that of the right table. If there had been two records in which an individual had the zip code of 13053 and age of 14, then there would have only been a $\frac{1}{2}$ chance of correctly guessing which record belonged to George Leary. This leads to the concept of an equivalence class:

Definition 2.5 (Equivalence class)**.** An equivalence class describes a set of records that contains identical values for their quasi-identifiers [10].

For example, in the previous example, records 1 and 4 belong to the same equivalence class because their values for the three quasi-identifiers are identical: zip code of 13051, age of 21, and nationality of US:

	Zip Code	Age	Nationality	Condition
1	13051	21	US	Heart Disease
4	13051	21	US	Viral Infection

The concept of k**-anonymity** relies upon the concept that, by controlling the size of equivalence classes, re-identification can be hindered. If, for example, the smallest equivalence class of a certain dataset contains 5 members, then in most cases an adversary with a linkable dataset will have at most a $\frac{1}{5}$ probability of linking a specific record with a record from an outside dataset, since each member of the equivalence class is identical in terms of its quasi-identifiers. The formal definition of k-anonymity follows:

Definition 2.6 (k-anonymity)**.** A dataset is k-anonymous if the size of the smallest equivalence class is equal to or greater than k.

Equivalently, k-anonymity can also be expressed as the requirement that the information released in a single record about a person is indistinguishable from at least k-1 distinct individuals in the dataset in terms of its quasi-identifiers [7]. For datasets that are required by law to satisfy k-anonymity before release, the value of k may often be determined by law or corporate policy, with a higher value of k corresponding to a stricter privacy standard.

	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	14857	$\geq$ 40	*	Viral Infection
6	14857	$\geq$ 40	*	Heart Disease
7	14857	$\geq$ 40	*	Viral Infection
8	14857	$\geq$ 40	*	Viral Infection
9	13060	3*	*	Cancer
10	13060	3*	*	Cancer
11	13060	3*	*	Cancer
12	13060	3*	*	Cancer

TABLE 2.3: This table is k-anonymous with $k = 4$ because each of the three equivalence classes (records 1-4, 5-8, and 9-12) contain at least 4 individuals that share the same values for all of their quasi-identifiers.

Although a k-anonymous dataset should theoretically ensure that the highest chance of re-identifying given record is $\frac{1}{k}$, a k-anonymous dataset does contain vulnerabilities in terms of re-identification, of which two are described below.

I. **Homogeneity attack.** If *all* members of an equivalence class have the same value for a given non-quasi-identifier attribute, then the value of that attribute becomes revealed for all members of that equivalence class. For example, in the above 4-anonymous table, all known individuals whose zip code is 13060 and are in their 30s are revealed to have cancer [8]. This means that, if I know that my neighbor is included in this dataset and that he lives in zip code 13060 and is in his 30s, the value of his sensitive field, cancer, has been revealed due to the fact that all members of his equivalence class have this value.

II. **Background knowledge attack.** The combination of background knowledge with a given dataset can also lead to violation of privacy in a k-anonymous table. In the above 4-anonymous table, imagine that one of the holders of this data knows that his Japanese friend, whose age is below 30 and who resides in zip code 130**, exists somewhere in this dataset. Since Japanese have extremely low rates of heart disease, then it is revealed with high probability that his friend has a viral infection, since that is the only other condition represented in his friend's equivalence class [8].

Other anonymization schemes exist, like l-diversity, that address these vulnerabilities, as discussed in the next section.

A modification to k-anonymization called (α, k)-anonymization provides further privacy protection in datasets where a subset of values within a certain attribute are sensitive. For example, a health dataset may contain a "condition" attribute that can take on the value of many common diseases (i.e., flu, cold, fever), as well as a few sensitive values (i.e., HIV or malaria) that need protection. In such a case, (α, k)-anonymity requires the dataset to meet the standards for k-anonymization, as well as meeting an additional requirement that, within each equivalence class, the relative frequency of any given positive sensitive value must be less than or equal to α. This acts as an upper bound for the confidence of determining a sensitive value given the value of the quasi-identifier.

For example, the below table is (α, k)-anonymous with $\alpha = 0.5$ and $k = 2$. We know that k=2 because there exist at least 2 records in each equivalence class (records 1 and 2; records 3 and 4). In order to determine the value of α, we must look at each equivalence

class separately. Within the first equivalence class, $\alpha_1 = \frac{1}{2} = 0.5$ because one of the two records is positive in its sensitive value (i.e., where the "condition" value is HIV). Within the second equivalence class, $\alpha_2 = \frac{0}{2} = 0$ because neither row is positive in its sensitive value. Therefore, overall, $\alpha = \max(0, 0.5) = 0.5$ and therefore $\alpha = 0.5$ for the whole dataset. Intuitively, this says that an adversary will never be more than 50% sure, for a given equivalence class, which sensitive value in a given equivalence class is positive [11]. Note that (α, k)-anonymization is only applicable in situations where only a subset of values within a certain attribute is considered to be "sensitive".

	Zip Code	Age	Condition
1	14857	35	HIV
2	14857	35	Flu
3	13060	25	Flu
4	13060	25	Flu

TABLE 2.4: This table is (α, k)-anonymous with $\alpha = 0.5$ and $k = 2$, because each equivalence class contains at least 2 records, and within each record there is at most a 0.5 chance of correctly guessing which record has a positive value for its sensitive attribute, HIV.

Throughout this paper, k-anonymity will be focused upon as the primary de-identification scheme, since it is the method of anonymization that was determined to be required by Family Educational Rights and Privacy Act (FERPA) law, which governs the release of the edX dataset that will be examined in this paper.

2.3.2 l-diversity

An alternative definition of de-identification, called **l-diversity**, addresses the vulnerabilities of k-anonymity by safeguarding against both homogeneity and background knowledge attacks. This de-identification scheme requires that each equivalence class must contain at least l distinct, well-represented values of any sensitive attribute, rather than simply requiring a certain number of rows within each equivalence class, as k-anonymity does.

Definition 2.7 (l-diversity)**.** A dataset is l-diverse if each equivalence class contains at least l "well-represented" values of any sensitive attribute [8].

"Well-represented" can be defined in multiple ways, such as requiring that a dataset contains enough information such that any given equivalence class does not reveal too much information about its subjects, or that the distributions of the sensitive values be relatively even. The ability of given researchers to set their own definitions of "well-represented" allows considerable flexibility in controlling what properties of a dataset

determine whether it is privacy-invasive in the context of l-diversity.

The notion of l-diversity where $l > 1$ prevents homogeneity attacks by requiring that records with the same quasi-identifiers not have *all* the same values in their sensitive fields. Background knowledge attacks are also protected against, by requiring any adversary to eliminate $l-1$ other features in order to deduce which sensitive value a given person has [8].

	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
2	1305*	≤ 40	*	Viral Infection
3	1305*	≤ 40	*	Cancer
4	1305*	≤ 40	*	Cancer
5	1485*	>40	*	Viral Infection
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Viral Infection
8	1485*	>40	*	Viral Infection
9	1306*	≤ 40	*	Heart Disease
10	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

TABLE 2.5: The above table is l-diverse with $l = 3$ because each group of individuals who share the same quasi-identifiers have at least 3 different values corresponding to the sensitive field, condition [8].

Extensions to l-diversity specify technical representations of what properties make a certain value "well-represented". One such extension is entropy l-diversity, defined as follows.

Definition 2.8 (Entropy l-diversity)**.** A dataset is Entropy l-diverse for a given choice of l if each equivalence class satisfies the condition that

$$-\sum_{s \in S} p_{(q*,s)} \log(p_{(q*,s)}) \geq \log(l)$$

where s is equal to the sensitive attribute, where $p_{(q*,s)} = \frac{n_{(q*,s)}}{\sum_{s' \in S} n_{(q*,s')}}$, and where $n_{(q*,s)}$ is the number of people with a given equivalence class $q*$ with sensitive value s and where S is the set of all possible values that the sensitive attribute can take.

Entropy l-diversity relies upon the concept that the *entropy* of a given dataset, defined by $-\sum_i P(x_i) \log(P(x_i))$, is an information metric for how much information is missing

– the larger the value, the more is missing. This follows intuitively from the mathematical definition. Consider the case in which there are two possible values for x_i. Then, if $P(x_1) = P(x_2) = 0.5$, then this means we are maximally uncertain about the outcome of x_i, and we see that this corresponds to a value of entropy of 1. Similarly, we see that entropy is minimized when we are certain about the values of x, i.e., when $P(x_1) = 1$ and $P(x_2) = 0$ or vice versa, which corresponds to an entropy of 0. Thus, the more distinguishable certain records are from each other, the lower the entropy.

Entropy l-diversity therefore sets a lower limit on how much information can be encoded, in order to ensure that every equivalence class does not reveal too much about any given record [8].

2.3.3 Differential privacy

Perhaps the "strongest" privacy requirement is that of *differential* privacy, which requires that the amount by which a dataset changes should be minimally affected by whether any specific individual is included or not. Therefore, the specific data of any given individual should never greatly affect a statistical analysis. More formally:

Definition 2.9 (ϵ-differential privacy)**.** A randomized function κ provides ϵ-differential privacy if, for all datasets x and x' that differ on at most one attribute, and all S $\subset$ Range(κ),

$$P(\kappa(x) \in S) \leq \exp(\epsilon) \cdot P(\kappa(x') \in S)$$

Intuitively, this requires that a function must perform similarly (i.e., multiplicatively within a constant of $\exp(\epsilon)$) on similar datasets to each other. This means that an individual's privacy risk should not substantially change whether they are included versus excluded in a differentially-private dataset. Differential privacy can be achieved by adding noise to each row with a Laplacian distribution proportional to the maximum difference between values of a single row [12].

2.4 How is de-identification achieved?

In order to transform a dataset from its original version to satisfy one of the above definitions of privacy, one of several techniques must be employed.

Often, the anonymization technique depends on which de-identification standard is being used. Differential privacy, for example, uses a technique of injecting "noise" to numeric attributes in the form of a random variable with a pre-determined mean and standard deviation in order to de-identify data.

For k-anonymization, which will be used throughout this paper, the two most utilized techniques are *generalization* and *suppression*. Generalization occurs when granular values are combined in order to create a broader category. This may occur for numerical variables (i.e., combining ages 20, 21,and 22 into a broader category of 20-22) or categorical variables (i.e., generalizing location-level data from "Boston" to a more general value, "Massachusetts"). Suppression, on the other hand, occurs when a record that violates anonymity standards is deleted from the dataset entirely [2].

There exists a delicate tradeoff between favoring generalization versus suppression during k-anonymization. Whereas a generalization-only de-identification approach prevents changes in the distribution of non-quasi-identifier fields by retaining all of their values in the dataset, the resulting values of quasi-identifiers may become generalized to a point where few conclusions can be made about their relationship with other fields. For example, if k-anonymity requires that the quasi-identifier "age" field in a given dataset be generalized from integer granularity to decade granularity, correlations between age and other characteristics would be difficult to calculate. Furthermore, since generalization is applied to a whole column, it decreases the quality of the *entire* dataset by broadening values for every record in a given dataset, whereas suppression only decreases the quality of the dataset on a record-by-record basis by excluding a given record.

A suppression-only de-identification approach, however, skews the integrity of a dataset. Although suppression only affects single records at a time, it alters the distribution of attributes by eliminating records from the dataset. If values are eliminated disproportionately to the original distribution of the data, this causes statistical bias in resulting data analyses. A balance must therefore be drawn between suppression and generalization.

The need for this balance between suppression and generalization in a dataset is reminiscent of the ubiquitous bias-variance tradeoff in statistics. This concept describes the tradeoff necessary when fitting models that are meant to generalize beyond the data on which they are trained. If a model is fit too closely to the training data, the model may not be generalizable to other samples due to its being overfit to the noise, rather

than the signal, of the training data. On the other hand, if the model is too loosely fit to the data in an attempt to be generalizable, it might miss important details in the relationship between variables.

This concept applies to the context of de-identification, where an anonymized dataset can be thought of as a "model" of the original dataset. If many records are suppressed, then the anonymized dataset may lack some of the intricacies in the relationship between variables, but if too many attributes are generalized, the variance of the conclusions that can be made from the anonymized dataset may be too high to motivate meaningful results.

Mean-squared error (MSE) is a useful metric that enables the ideal balance between bias and variance to be quantified. The MSE of a given model is equal to the mean value of the difference between the model's predicted value for a certain outcome and the actual outcome, which can mathematically proven to be equal to the bias^2 + variance of a given model. Therefore, since minimizing MSE requires a balance between a low bias and variance, the MSE may be a useful numeric measure of how much "error" is being introduced into the dataset at any given point in the de-identification process. This may be useful, for example, in order to decide whether to generalize attributes versus suppress records during a certain step of the de-identification process.

2.5 Determination of the "best" dataset for release

During the k-anonymization process, even a single choice between row suppression and column generalization can cascade to produce two largely different de-identified datasets that both satisfy k-anonymity. Thus, it follows that a large number of different datasets can satisfy the requirements for any given de-identification standard. Given that the choice of which of these datasets to publish can have a large bearing on the analysis that follows, is there a way to generate the dataset that is "best" for public release?

If the statistical analyses that will be performed on the dataset are known beforehand, then the de-identification process can be executed in order to minimize the error of that analysis. Previous literature has illustrated how alterations to de-identification schemes can optimize for specific post-release data analyses, and are described below.

2.5.1 Count Data

In many cases, a researcher may be concerned about preserving the marginal distribution of a single attribute. In a medical dataset, for example, preserving counts of patient diagnoses may be important in understanding the relative frequencies of the occurrence of certain diseases.

Count data of all non-quasi-identifier attributes can easily be maintained by only using generalization, rather than suppression, of records. Generalization preserves count data by nature of its not excluding any records. In the case where the counts of a quasi-identifier attribute are desired, a generalization approach will still maintain counts accurately, although with reduced granularity of the values. For example, if zip code is generalized to city names, there will exist no ability for researchers to recover counts at a level more granular than the city level.

Swapping of quasi-identifier values is another count-preserving method that can be used during the k-anonymity process. This technique preserves the existence of quasi-identifier values within the dataset, simply altering the association to other attributes. Consider the below table, where state and gender are the two quasi-identifying fields. The table is 1-anonymous because each equivalence class contains one record – every record has a unique combination of values for the state and gender variables.

State	Gender	Condition
MA	F	Flu
MA	M	Flu
NY	F	Cancer
NY	M	Cancer

If the dataset were required to be 2-anonymous, this could be achieved via a generalization of MA and NY into a broader category like “Eastern US”. However, this sacrifices the granularity of knowledge about states.

Alternatively, quasi-identifier values could be swapped in order to achieve k-anonymity. By switching the gender of the second and third record as shown in the below right table, this dataset can be made 2-anonymous without sacrificing the current granularity level of the dataset nor the counts of either one of the quasi-identifier fields. Although this swapping method preserves the granularity of quasi-identifiers, it does so at the cost of introducing a false association between genders of the second and third rows with their respective conditions.

State	Gender	Condition
MA	F	Flu
MA	M	Flu
NY	F	Cancer
NY	M	Cancer

→

State	Gender	Condition
MA	F	Flu
MA	F	Flu
NY	M	Cancer
NY	M	Cancer

TABLE 2.6: This table illustrates the de-identification of the left table from its 1-anonymous to its 2-anonymous version by swapping quasi-identifier values between rows, in order to preserve their frequencies. This method, however, misrepresents the association between the swapped quasi-identifiers with their other attributes.

The above-described methods that preserve marginal distributions of single attributes may therefore introduce inaccuracies in the *joint* distributions between multiple attributes. For example, while the percentage of people within a dataset who have a certain medical condition may be preserved, the relationship of the correlation of the "medical condition" variable with demographic or lifestyle attributes may be significantly altered, which is often more important to researchers (i.e., how many percentage of people who have a certain condition are also smokers?).

2.5.2 Clustering

Clustering is an unsupervised learning algorithm that attempts to find a pre-specified number of groups n, in a dataset, such that the members within each group are more similar to each other than to members of other groups.

Fung et. al. discuss a framework in order to find the most optimal k-anonymous dataset that preserves cluster analysis for a prespecified number of clusters, n. This algorithm works by clustering the original, unanonymized dataset and labeling each datapoint with its corresponding cluster assignment, and then de-identifying the dataset in a top-down measure that preserves the original cluster assignments as much as possible with each generalization step. This top-down method is called *top-down refinement*, in which researchers pre-specify different levels of generalization for each distinct value in a dataset.

For example, a researcher might pre-specify that a certain numeric attribute may be generalized into bins of size 1, 2, 5, or 10. Data is then initialized at the most masked or generalized values (i.e., a bin size of 10), and then is iteratively refined into more specific values until it can no longer be refined without breaking the k-anonymity requirement. The resulting cluster analysis is empirically shown to perform 24% to 125% better than a naively chosen k-anonymous dataset [6].

One disadvantage of this approach as it relates to clustering, however, is the inability to know beforehand how many clusters a certain researcher will be interested in using. To address this, Fung et al. propose the release of a different dataset for each value of n. The risk of re-identification due to the release of multiple datasets is not discussed in their paper.

This approach is valuable in its generalizability to almost any statistical analyses – one promising approach to minimize modeling error introduced by de-identification is to simply perform the statistical analysis beforehand and then de-identify the dataset in order to minimize the error of that task at every generalization step, as was done in this procedure.

2.5.3 Linear regression

LeFevre et. al. discuss a "greedy" mechanism that optimizes the de-identification procedure for linear regression tasks. At each step, a top-down refinement of the dataset is performed that minimizes the following expression, where m is the number of datapoints, T is the dependent variable being modeled, and $V_1, ..., V_m$ denotes the set of data partitions resulting from a certain candidate split:

$$\sum_{i=1}^{m} \sum_{t \in V_i} (t.T - \bar{T}(V_i))^2$$

We note that the above equation represents the squared difference between the linear regression's prediction of a given value and the true mean of that value. Therefore, minimizing this value intuitively creates a dataset whose resulting linear regression will have its accuracy maximized. This algorithm is repeated until no further refinements can be made that do not violate the given anonymity requirement. Experimental results showed that this algorithm generally led to models with lower mean absolute errors than naive de-identification procedures, for all values of k in k-anonymization [5].

2.5.4 Classification

Classification is a supervised learning algorithm that attempts to use characteristic features of data points in order to predict which of two (or more) categories a specific data point belongs to. Classification methods may, for example, be used in order to build a predictive model for which disease a patient is likely to have, based on their demographic attributes. De-identification procedures that are optimized for two popular classification methods, support vector machines and logistic regression, are discussed below.

Support Vector Machines

The support vector machine (SVM) is a supervised classification algorithm that uses labelled training data in order to determine a classification boundary that optimally divides data into its two labelled classes. The accuracy of SVMs are particularly affected by data de-identification because, if records that are near the classification boundary become generalized or suppressed, the classification boundary becomes displaced significantly from its "true" value in the original dataset, therefore dramatically reducing the performance of the classifier.

In addition, any *non-linear* SVM classifier that is trained on the original data cannot be released to the public because the classifier itself contains information about *support vectors*, the data samples that lie closest to the decision boundary [13]. To this end, Keng-Pei Lin and Ming-Syan Chen describe the mechanisms behind a Privacy-Preserving SVM classifier (PPSVC) that is able to accurately train the classifier without disclosing sensitive personally identifying information. The PPSVC works by transforming the sensitive attribute values of support vectors using a Taylor polynomial of linear combinations of monomial feature-mapped support vectors, thereby *approximating* the decision function of the SVM while maintaining a comparable level of accuracy [13]. Although the Taylor series is an infinite series, a relatively accurate approximation can be easily found by only using lower-order terms. Under this model, the sensitive content of any support vector cannot be recovered without knowing the values of every other support vector (since the coefficients of the model are a linear combination of all of the support vectors).

Logistic Regression

Logistic regression is a technique that uses labelled training data in order to build a classifier that predicts the probability that a given record belongs to a certain class, given

certain attributes about the record. For example, in a dataset that contains demographic information of patients and whether they have the flu or not, a logistic regression might be used in order to generate predictions of the probability that a new patient, given his or her demographic information, will have the flu.

LeFevre et. al. develop an optimal partitioning mechanism, in which a top-down refinement of data is performed that minimizes the entropy across partitions, where P is the current tuple of quasi-identifiers, P' denotes the set of partitions resulting from a given candidate split, and $p(c|P')$ is the proportion of tuples in P' that are labeled with class $C = c$, where D_C is the set of possible class labels [5]:

$$\sum_{P'} \frac{|P'|}{|P|} \sum_{c \in D_C} -p(c|P') \log(p(c|P'))$$

Notably, the second half of this equation, $\sum_{c \in D_C} -p(c|P') \log(p(c|P')$, is equal to the entropy of a dataset, which can be thought of as the amount of information that is missing from a dataset. The higher the entropy, the more information is missing from the dataset. Therefore, the intuition behind this optimal partitioning mechanism is to minimize the weighted entropy of data within each class, therefore maximally preserving the amount of information within each class.

2.6 Unspecialized datasets for release

Though the above sections provide a theoretical framework for de-identifying datasets that are predetermined to be likely used for a certain research purpose, research objectives of a given dataset are often unpredictable. Releasing multiple datasets that are optimized for several different data analyses is problematic, because the aggregation of all released datasets must collectively satisfy de-identification standards, which may decrease the quality of each individual dataset.

Furthermore, especially for high-quality datasets that contain many attributes and may be accessed publicly by many researchers, de-identification of a dataset for a predetermined research purpose may actually be *un*desirable in that the preservation of relationships between certain variables may cause distortions in relationships between other variables. None of the literature reviewed in the above section specifically addressed this potential effect.

For these reasons, there are many merits for a general approach that can de-identify datasets without optimizing for a specific statistical analysis. This paper will thus examine the aspects of the de-identification processes that preserve *general* properties of datasets, such as distributions of numeric attributes, rather than de-identifying datasets optimized for specific data analyses.

Chapter 3

Experimental approach

3.1 Dataset description

edX is a massive online open course (MOOC) provider jointly founded by Harvard University and the Massachusetts Institute of Technology in May 2012 that offers university-level classes to the general public, often free of charge. A dataset containing student records from sixteen courses on the edX platform was de-identified and released to the public on May 30, 2014. It provides a valuable case study into the differences that are introduced by the de-identification of a dataset.

Before release, edX data was altered in order to comply with the Family Educational Rights and Privacy Act (FERPA). FERPA requires that personally identifiable information be removed, as defined by fields like name, address, social security number, and mother's maiden name. However, it also requires that other information, alone or in combination, must not enable identification of any student with "reasonable certainty" (34 C.F.R. § 99.3,2013).

In order to satisfy the latter clause, the edX research team opted for a k-anonymity framework. The value of k was chosen to be 5 due to the U.S. Department of Education's Privacy Technical Assistance Center which claimed that "statisticians consider a cell size of 3 to be the absolute minimum" and that values of 5 to 10 are even safer [2].

This paper analyzes the subset of the data corresponding to HarvardX (not MITx) courses. De-identification was performed on this HarvardX-only dataset to produce a

5-anonymous version that is referred to as the "de-identified" dataset.

Attribute	Distinct values: Original dataset	Distinct values: De-identified dataset	Identifier	Quasi-identifier
User ID	363425	312297		
Course ID	5	5		Yes
Username	363412	N/A	Yes	
Registered	1	1		
Viewed	2	2		
Explored	2	2		
Certified	2	2		
IP	211641	N/A	Yes	
Country name by IP	226	N/A		
Country by address	207	N/A		
Level of education	14	7		Yes
Year of birth	115	79		Yes
Gender	5	4		Yes
Grade	103	103		
Start time	413	413		
Last event	388	388		
Number of course interactions	9238	5627		
Modal IP	210322	N/A		
Modal IP country	226	N/A		
Final country	229	22		Yes
Final country source	4	N/A		
Number of active days	182	135		
Number of video plays	3322	2045		
Number of chapters	35	35		
Number of forum posts	183	8		Yes
Roles	1	1		

TABLE 3.1: Description of original and de-identified datasets containing only HarvardX courses.

The dataset that was studied contained 26 attributes that contain information on the demographics, activity levels, and performance levels for each student-course combination included in the dataset. The quasi-identifiers are considered to be course ID, level of education, year of birth, gender, country, and the number of forum posts. The number of forum posts is considered to be a quasi-identifier because the forum was a publicly accessible website that could be scraped in order to link user IDs with their number of forum posts. Course ID was also considered to be a quasi-identifier because unique combinations of courses could conceivably provide a link between personally identifiable information that a student posts in a forum with the edX dataset.

3.2 De-identification procedure for edX dataset

De-identification of the edX dataset was achieved via a Python script written by Jon Daries, a senior research analyst at MIT Institutional Research [14]. It is written in order to allow the user to control the degree of generalization, suppression, and numerical binning that is performed, as well as allowing different values of k for k-anonymization

to be set.

At a high level, the algorithm works by loading in two versions of a SQLite database containing the original data – one that will be modified by the de-identification process and the other of which is preserved for comparison with the de-identified dataset. The first database is then sanitized by removing sensitive data such as timestamps, and then relevant attributes are generalized, suppressed, trimmed, and binned. Finally, records that violate k-anonymity after these four steps become suppressed in order to result in the final k-anonymous dataset [14].

This algorithm, based on Latanya Sweeney's Datafly algorithm, with implementation specific to the edX dataset, is characterized by its use of bottom-up generalization rather than top-down specialization, meaning that discrete values start at their original values and are generalized to broader values as necessitated by the k-anonymization. Furthermore, Datafly is greedy in that each step is chosen in terms of the most locally (rather than globally) optimal step [15].

Below is a detailed outline of the algorithm used to de-identify the edX dataset.

I. Define environment variables.
   The user assigns string values to the following variables:
   - `file`, the original non-de-identified dataset filename
   - `table`, the name for the SQLite database that will be modified
   - `userVar`, the column name of the user ID attribute
   - `courseVar`, the column name of the course ID attribute
   - `countryVar`, the column name of the country name attribute
   - `k`, the level of k-anonymity that is required

II. Load data into SQLite database.
   The original data is loaded in twice as two separate SQLite databases: the first will ultimately undergo the de-identification procedure, and the second will serve as an unaltered database that will be used for comparisons with the de-identified dataset at the end.

III. The original dataset undergoes basic sanitation.
   A. Anonymized versions of potentially identifying variables are created.

i. The timestamp of the date of course registration and the date of the last interaction with the course are removed, leaving only a date in the form of `YYYY-MM-DD`).

ii. New attributes are created that correspond to the full country name and full region name (i.e., "India" and "Other South Asia") to which each record's country code corresponds.

B. Indexes are created on the `courseVar` and `userVar` variables to optimize searches and queries on these columns. (Users cannot see these indices, however.)

C. All staff and instructors are deleted from the table.

D. Anonymous user IDs are generated, which take the form of "MHxPC13" + random number, where the random number is generated using a salted hash of the username (i.e., a hash of the username concatenated with a random number, which makes it harder for the hash to be reversed). The choice of "MHxPC13" is a shortened version of "MITx/HarvardX Person-Course AY2013".

IV. Calculate entropy.

A. Create a new attribute called `entropy` that contains concatenated versions of the values of every column.

B. The entropy of the dataset is then calculated. If the set of unique concatenated values is viewed as a random variable X with possible values $\{x_1, ..., x_n\}$, and if the probability mass function $P(x_i)$ is defined as the number of times that the data takes on the value of x_i over the total number of rows there are in the data, then entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2[P(x_i)]$$

C. The variables that are indicative of course performance (`viewed`, `explored`, `certified`, `grade`, `nevents`, `ndays_act`,`nplay_video`, `nchapters`, `nforum_posts`) are loaded into the `utilVars` variable and then passed through the `utilMatrix` function in order to calculate the entropy, mean, and standard of these variables. This allows a record to be kept of how much these course performance variables change throughout the de-identification process.

V. Generalize and drop variables as needed.

A. Every unique course combination is found. Users who have taken a course combination whose value is taken by less than $k-1$ other users then must have

a given number of courses dropped until their course combination is no longer unique. Users with course combinations that cannot be made non-unique by this method are dropped from the dataset.

B. For rows whose "country" values are represented less than k times, replace the "country" with the "continent" value.

VI. Check for k-anonymity.
At this point, the program prompts the user for the indices of the quasi-identifying (QI) variables and then reports how many rows do *not* satisfy k-anonymity given those QI fields. The program also checks k-anonymity for rows that contain null values by checking whether the unique combinations of the non-null values of the QI variables satisfied k-anonymity.

VII. Data trimming and binning.

A. The `tailFinder` function is then applied to the two numeric quasi-identifier fields: `nforum_posts` and `YoB` (year of birth). The `tailFinder` function cuts off user-specified values on the low or high ends of the values for a given column. This can be helpful for eliminating unrealistic values (i.e., birth years before 1900), values that are clearly too identifying (i.e., forum posts above 200), or simply trimming the range of values down to an evenly-divisible number, which makes binning numbers into integer-sized buckets more cleanly performed.

B. After cutting the tails, the values for these variables are binned into a user-specified number of bins and then the `kAnonWrap` function is used to check for k-anonymity and to return the proportion of rows required in order for the dataset to be k-anonymous.

VIII. Data suppression check.
Finally, after data trimming and binning, each record's QI attributes are concatenated in order to form a QI key. Rows whose QI key are represented less than k times are dropped from the dataset in order to generate the final, de-identified dataset.

IX. Comparison between original and de-identified datasets.
After the de-identified dataset has been generated, basic statistics regarding the total counts and averages of demographic and course performance attributes are reported, for both the original and de-identified datasets, as a simple measure of the bias introduced by the de-identification process [14].

3.3 edX data: de-identified versus original dataset

Here we investigate the differences between the original edX data and the k-anonymized data that were introduced during the de-identification algorithm described above. By FERPA requirements, k was set to 5, meaning that there can exist no student record that is indistinguishable from less than 4 other student records in terms of its quasi-identifying fields: course ID, gender, year of birth, country, level of education, and number of forum posts [16].

The de-identification process decreased the number of records in the dataset from 440853 to 340354, a 22.8% reduction. Here we investigate in more detail how the characteristics of several attributes were changed between the original and de-identified dataset.

3.4 One-dimensional changes: count data

During the process of k-anonymization, the generalization and suppression of records leads to changes in the summary statistics of attributes. In the edX dataset, the count data of demographic and enrollment information changed as depicted in the tables below as a result of the de-identification process:

	Original	De-identified	% Decrease
CB22x,S13	45577	30002	34.2%
CS50x,12	193495	169621	12.3%
ER22x,S13	81378	57406	29.4%
PH207x,F12	66145	41592	37.1%
PH278x,S13	54258	39602	27.0%

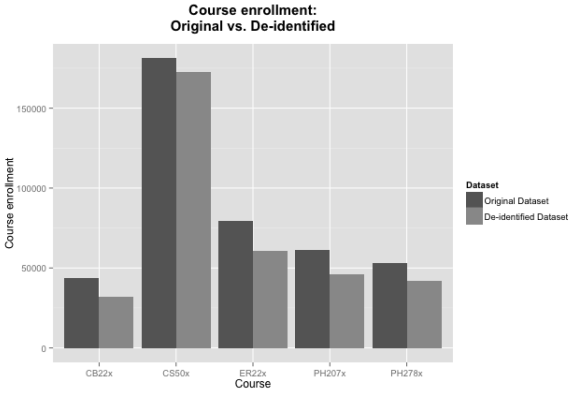


FIGURE 3.1: Changes in the enrollment of each class caused by the de-identification procedure.

	Original	De-identified	% Decrease
Male	0.60	0.62	-3.33%
Viewed	0.57	0.57	0%
Explored	0.80	0.60	25%
Certified	0.40	0.20	50%

TABLE 3.2: Changes in performance and gender variables caused by the de-identification procedure.

We also note that the distribution of numeric attributes changes between the original and the de-identified datasets. Here, we calculate the *chi-squared distance* between attributes in the original versus the de-identified datasets. This metric measures the "distance" between the distributions of two variables, as defined by the following equation, where `orig` and `anon` are two vectors containing *normalized* counts of binned values of a given attribute in the original and de-identified datasets:

$$d(\text{orig, anon}) = \sum_{i=1}^{n} \frac{(orig_i - anon_i)^2}{2(orig_i + anon_i)}$$

Intuitively, this equation measures how much larger the observed differences between the distribution of data between the original and anonymized datasets (the numerator) are from the sum of the values across the data (the denominator). A larger value of the chi-squared distance indicates a larger difference between the two distributions of data.

Attribute	**Bin size**	**Chi-squared distance**
Number of video plays	1	0.034
Number of forum posts	1	0.020
Number of active days	1	0.0083
Number of chapters accessed	1	0.0080
Year of Birth	1	0.0064
Grade	0.01	0.0016

TABLE 3.3: Difference in the distribution of attributes between the original and anonymized datasets, as measured by chi-squared distance. A higher value indicates a larger difference between the two distributions of data.

This table reveals that the largest differences between the original and de-identified datasets are found in the number of forum posts and the number of video plays, both of which are attributes measuring student activity. This suggests that the de-identified dataset tended to be skewed in its representation of how active students were.

The histograms on the next page visually characterize the differences between the original and de-identified datasets. Again, we observe the largest difference between the original and de-identified distributions of the number of forum posts, due to the suppression of rare values of forum posts during the k-anonymization process. We also generally see that other attributes measuring course activity (i.e., number of chapters accessed, number of active days, and number of video plays) appear to have their high values underrepresented in the de-identified dataset, likely also induced by the suppression of rows corresponding to more active students.

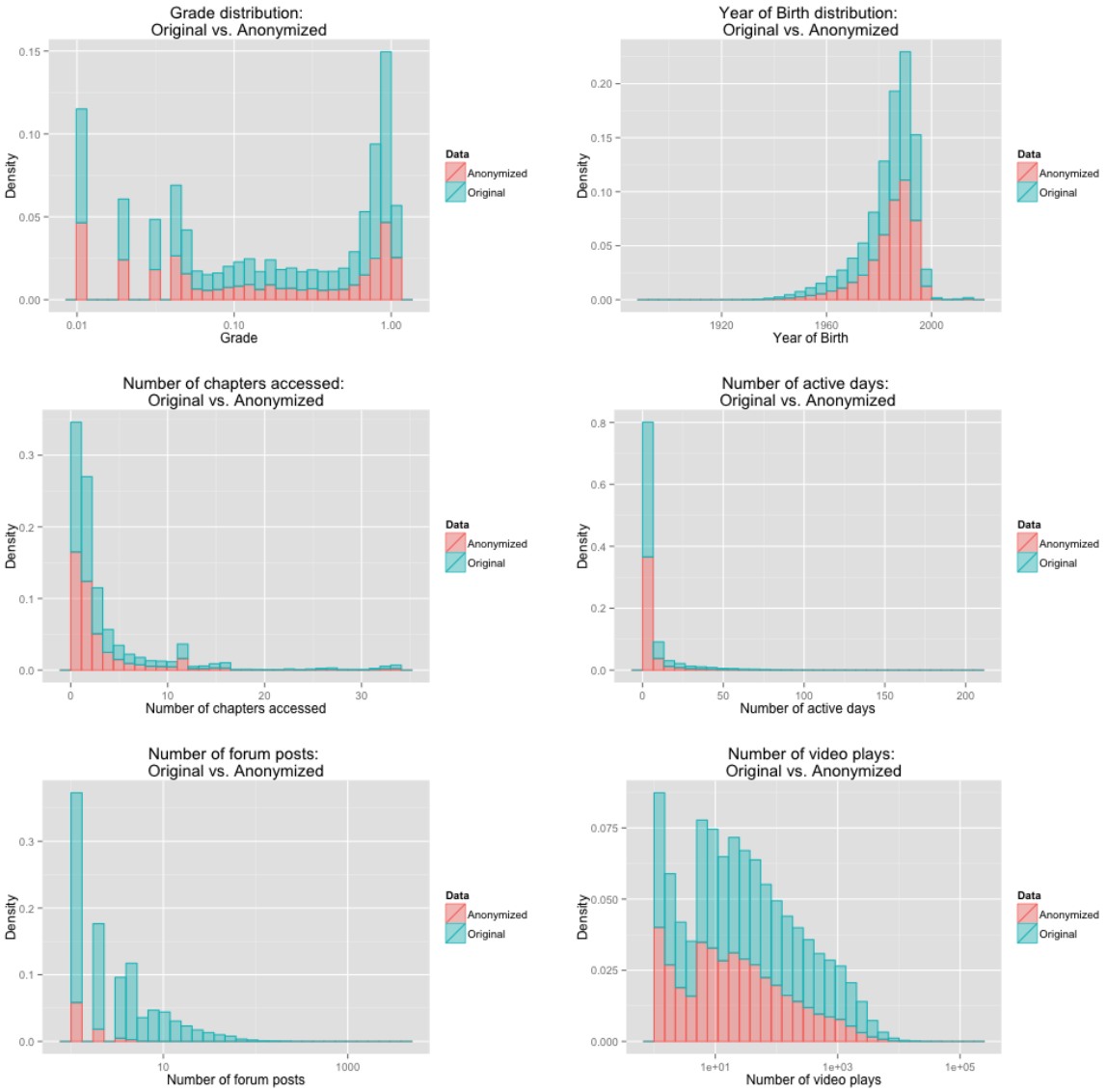


FIGURE 3.2: Histograms describing the differences between the original and de-identified dataset. The blue bars represent the frequency of values in the original dataset, whereas the red bars represent the frequency of values in the de-identified dataset. Most notably, the de-identified dataset often does not contain values in the tail ends of many activity variables, such as the number of forum posts and the number of active days, likely induced by the suppression of rows during the k-anonymization process.

3.5 Two-dimensional changes: correlation data

Correlation is a measure of the degree to which two variables have a linear relationship. A measure of 1 indicates a total positive linear relationship, -1 indicates a total negative linear relationship, and 0 indicates no linear relationship. Mathematically, correlation is defined as below.

Definition 3.1 (Correlation)**.** The correlation between two variables, X and Y, is given by the covariance of the variables divided by the product of their standard deviations:

$$Correlation(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_x \sigma_y}$$

Pairwise correlations between numeric attributes of the original and de-identified datasets are reported below:

	viewed	explored	certified	grade	ndays_act	nplay_video	nchapters	nforum_posts
viewed	1.000	0.250	0.163	0.195	0.242	0.074	0.413	0.046
explored	0.250	1.000	0.619	0.681	0.637	0.241	0.789	0.130
certified	0.163	0.619	1.000	0.940	0.658	0.251	0.642	0.156
grade	0.195	0.681	0.940	1.000	0.721	0.303	0.692	0.163
ndays_act	0.242	0.637	0.658	0.721	1.000	0.371	0.673	0.243
nplay_video	0.074	0.241	0.251	0.303	0.371	1.000	0.175	0.088
nchapters	0.413	0.789	0.642	0.692	0.673	0.175	1.000	0.163
nforum_posts	0.046	0.130	0.156	0.163	0.243	0.088	0.163	1.000

TABLE 3.4: The pairwise correlations for the **original** dataset between numeric attributes.

	viewed	explored	certified	grade	ndays_act	nplay_video	nchapters	nforum_posts
viewed	1.000	0.223	0.126	0.154	0.218	0.077	0.426	0.054
explored	0.223	1.000	0.518	0.574	0.552	0.243	0.765	0.034
certified	0.126	0.518	1.000	0.939	0.571	0.264	0.584	0.037
grade	0.154	0.574	0.939	1.000	0.634	0.322	0.626	0.044
ndays_act	0.218	0.552	0.571	0.634	1.000	0.419	0.616	0.057
nplay_video	0.077	0.243	0.264	0.322	0.419	1.000	0.200	0.003
nchapters	0.426	0.765	0.584	0.626	0.616	0.200	1.000	0.090
nforum_posts	0.054	0.034	0.037	0.044	0.057	0.003	0.090	1.000

TABLE 3.5: The pairwise correlations for the **anonymized** dataset between numeric attributes.

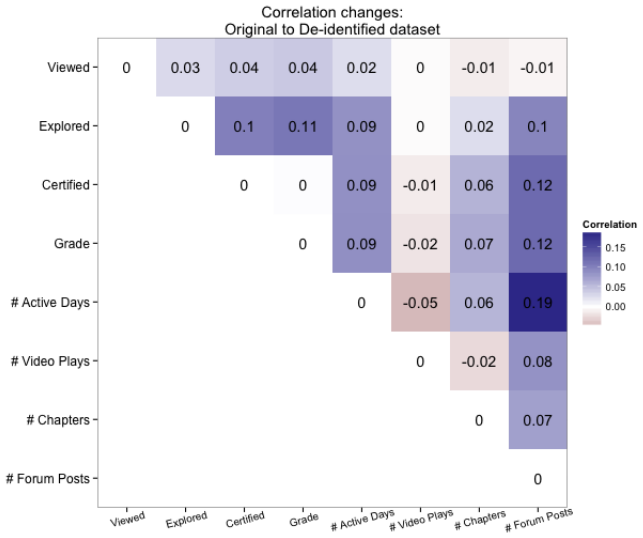


FIGURE 3.3: Matrix representing the **change** in correlation of numeric variables between the original and de-identified datasets.

The above figure shows the changes in correlation between the original and de-identified datasets. The largest differences between correlations in the original versus de-identified datasets occur in the rightmost column between the number of forum posts with other attributes.

The large changes in these correlations are likely caused by the fact that the number of forum posts is a quasi-identifier, meaning that uncommon values (which in this case corresponded to larger values) of this attribute correspond to records that must be suppressed. The suppression of records in the original dataset who posted very frequently on the forum (including one record with more than 3,500 forum posts!) but who had moderate levels of activity or performance therefore caused an increase in correlation of the de-identified dataset as compared with the original dataset.

There did exist other large changes in correlation, however, that were not necessarily associated with the elimination of outliers. For example, the correlation between grade and whether someone had accessed at least half of the chapters ("explored") had a change of 0.11, despite both variables being bounded between 0 and 1. Similarly, the correlation between whether someone became certified and whether someone had accessed at least half of the chapters ("explored") experienced a correlation change of 0.10, despite both

variables also only taking on values of either 0 or 1. Therefore, relationships between other course activity variables were affected by de-identification, as well.

Chapter 4

k-anonymization: identifying sources of bias

In order to better understand how the k-anonymization process creates undesirable differences between the original and de-identified datasets, we here focus on understanding the mechanisms that cause uneven suppression of records. As discussed in the previous chapter, the uniqueness of each record's combination of quasi-identifier values determines its likelihood of requiring suppression or generalization under a k-anonymization framework. The below section will further investigate this idea in order to determine how properties of quasi-identifiers within a dataset influence which records become generalized or suppressed.

4.1 The relationship between quasi-identifier frequency and the bias of attributes

During the de-identification process, the rarity of each record's combination of quasi-identifier attributes determines the likelihood of that record becoming suppressed or generalized. Therefore, in order to understand the mechanisms by which bias is introduced into a de-identified dataset, we first aim to understand how quasi-identifier combinations affect the skewness of the data created by the de-identification process.

We begin by investigating how the frequency of each quasi-identifier field in isolation maps onto non-sensitive column values like grade, performance variables, and activity variables. Intuitively, we hypothesize that highly negative correlations between the *frequency* of a given quasi-identifier value with other numeric attributes will generate

the most highly biased numeric variables in the de-identified dataset. For example, if rare values of a given quasi-identifier tend to be associated with high grades, de-identification would likely skew grades downward due to the suppression of records with high grades.

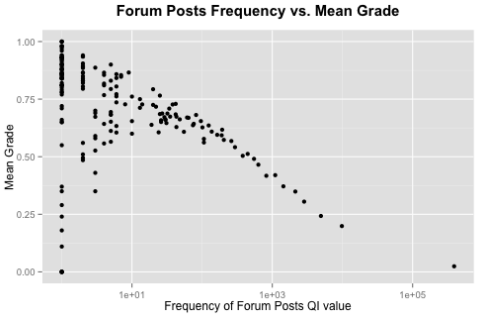
4.1.1 Individual quasi-identifier frequencies

Quasi-identifier frequencies versus grade

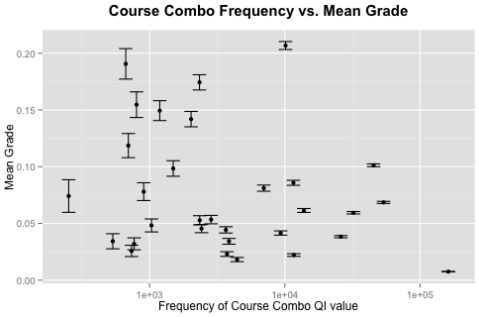
Of the six quasi-identifier fields, the number of forum posts has the strongest relationship between the frequency of its values with their corresponding mean grade. The negative sign of the correlation signals that, since rarer values of the number of forum posts tend to be associated with higher grades, the k-anonymization process is likely to skew grades downward by virtue of its suppression of records that correspond to rarer quasi-identifier combinations.

The rarity of a given user's course combination also appears to have a fairly strong negative correlation with mean grade, signaling that the exclusion of records based on this criteria may similarly tend to skew grades towards lower achievers.

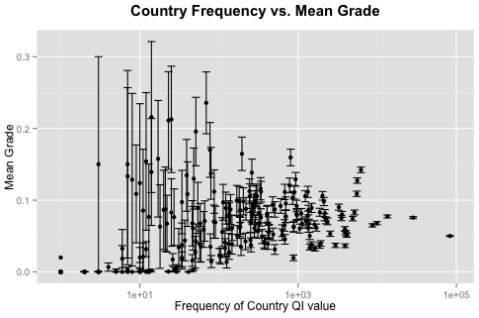
Interestingly, all six of the quasi-identifier fields show some degree of *negative* correlations between the frequency of the quasi-identifier value with grade, implying that k-anonymization would likely create a downward skew in grades. All graphs (on the next page) demonstrate the relationship between the frequency of quasi-identifier values versus the mean grade, and are shown for k=1, the original dataset.



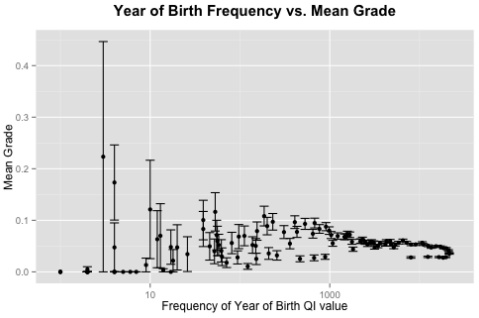
Correlation: -0.4283; p-value: $<$ 1e-15



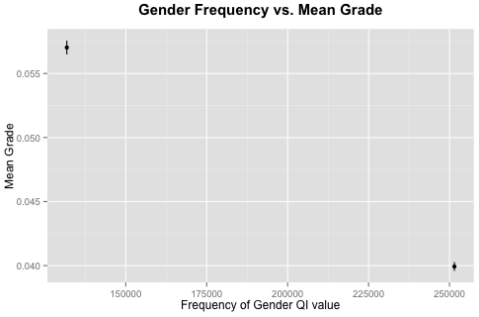
Correlation: -0.1698; p-value: $<$ 1e-15



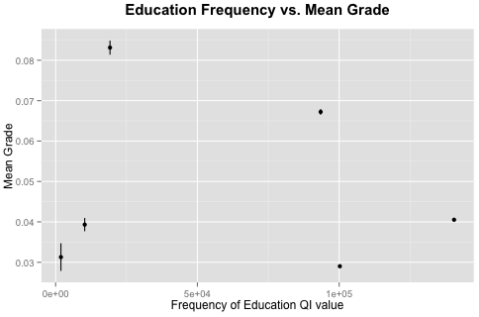
Correlation: -0.0512; p-value: $<$ 1e-15



Correlation: -0.0452; p-value: $<$ 1e-15



Correlation: -0.0447; p-value: $<$ 1e-15



Correlation: -0.0228; p-value: $<$ 1e-15

FIGURE 4.1: Here we explore the relationships between the frequency of each of the six quasi-identifier attributes with the corresponding mean grade for that value. Both visually and numerically, the strongest relationship clearly exists between the frequency of forum post values with the mean grade of those records.

Quasi-identifier frequencies versus performance variables

The number of forum posts is again the quasi-identifier variable with the clearest relationship between the frequency of the occurrence of its values with performance variables. From the fairly high negative correlations, we deduce that rarer values of the number of forum posts tend to be associated with people who view, explore, and are certified by the course. Therefore, this suggests that the k-anonymization process is likely to skew performance metrics downward, since these "rarer" values of the number of forum

posts are more likely to distinguish a given record in terms of its combination of quasi-identifier values.

The strong negative correlations between the frequency of the number of forum posts and the percent of students who viewed, explored, or were certified in a course is explained by the fact that most students post a low number of times (40 times more students posted 0 times than 1 time!), and it is precisely these inactive students who are most likely to *not* have viewed, explored, or been certified in a course. Similarly, people who post multiple times (which represent less frequently represented values of the number of forum posts) are the most likely to have viewed, explored, or been certified in a course. Thus, this results in a fairly strong negative relationship between the frequency of the value of the number of forum posts with certain performance metrics.

We also note that, as before, the vast majority of correlations between the frequency of quasi-identifier values and performance variables in the original dataset are negative, as shown in the table below. Again, this suggests that anonymization is likely to skew the performance metrics downward by removing more uncommon values of quasi-identifier fields that tend to be associated with higher performance values.

Correlations(QI freq, performance)			
QI var	Performance var	Correlation	p-value
Number of forum posts	Viewed	-0.2122	< 1e-15
	Explored	-0.3498	< 1e-15
	Certified	-0.3878	< 1e-15
Year of birth	Viewed	-0.0502	< 1e-15
	Explored	-0.0476	< 1e-15
	Certified	-0.0441	< 1e-15
Education	Viewed	-0.0317	< 1e-15
	Explored	-0.0291	< 1e-15
	Certified	-0.0208	< 1e-15
Country	Viewed	-0.0017*	0.3828
	Explored	-0.0469	< 1e-15
	Certified	-0.0420	< 1e-15
Course combinations	Viewed	0.0475	< 1e-15
	Explored	-0.0849	< 1e-15
	Certified	-0.1414	< 1e-15
Gender	Viewed	0.0304	< 1e-15
	Explored	0.0009*	0.581
	Certified	-0.0334	< 1e-15

TABLE 4.1: The numeric values of the correlation of quasi-identifier frequencies with various performance metrics in the original dataset. Correlation values that are marked with an asterisk are not significant at the $\alpha = 0.05$ level.

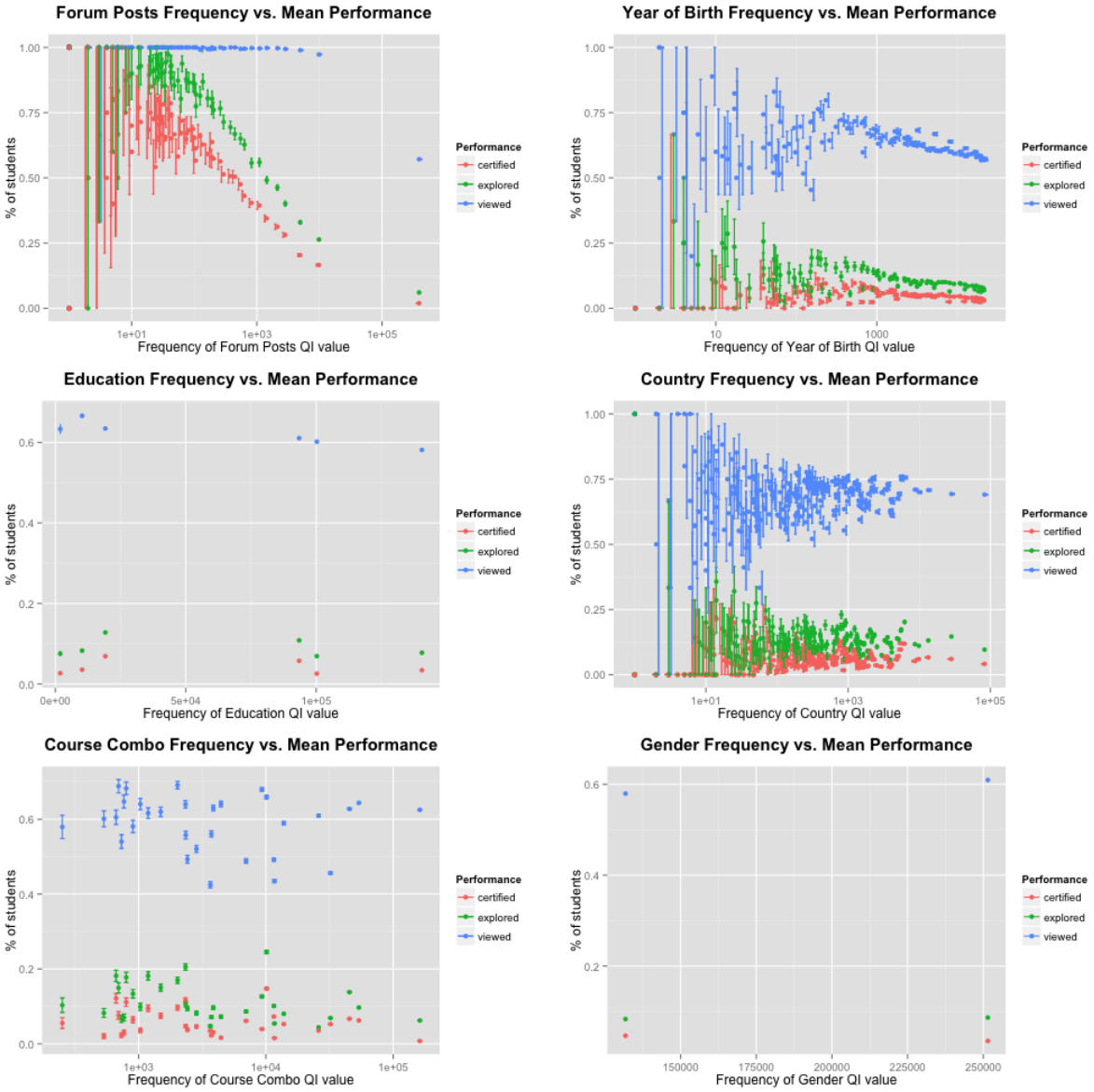


FIGURE 4.2: These graphs show the relationship between the frequency of each of the six quasi-identifier attributes with the mean percentage of students who viewed the courseware tab, accessed at least half the chapters ("explored") or earned a certificate in the course. Again, the strongest relationship occurs between the frequency of forum posts with each of these values.

Quasi-identifier frequencies versus activity variables

As we have seen above with the grade and performance variables, the number of forum posts again has the strongest association between its frequency of occurrence with activity variables, such as the number of chapters accessed, the number of active days on edX, the number of interactions (events) with the site, and the number of times videos were played. This follows the same intuition as above – the most frequently occurring value for the number of forum posts is 0, which is precisely the population of students

who are most likely to have not been active in a course. We also see a fairly strong negative correlation between these activity measures with two other quasi-identifier fields: frequency of course combinations and year of birth.

As before, the negative sign of all of these correlations suggests that the records that are most likely to be dropped during the k-anonymization process correspond to users with high activity levels. Therefore, the activity levels in the anonymized datasets are likely to be biased downward.

Correlations(QI freq, activity)			
QI var	Activity var	Correlation	p-value
Number of forum posts	# Chapters	-0.4264	< 1e-15
	# Active Days	-0.4736	< 1e-15
	# Events	-0.4124	< 1e-15
	# Video Plays	-0.2580	< 1e-15
Course combinations	# Chapters	-0.2601	< 1e-15
	# Active Days	-0.1152	< 1e-15
	# Events	-0.1358	< 1e-15
	# Video Plays	-0.0171	0.000259
Year of birth	# Chapters	-0.0825	< 1e-15
	# Active Days	-0.1014	< 1e-15
	# Events	-0.0713	< 1e-15
	# Video Plays	-0.0607	< 1e-15
Country	# Chapters	-0.0457	< 1e-15
	# Active Days	-0.0318	< 1e-15
	# Events	-0.0407	< 1e-15
	# Video Plays	-0.0399	< 1e-15
Gender	# Chapters	-0.0460	< 1e-15
	# Active Days	-0.0248	< 1e-15
	# Events	-0.0539	< 1e-15
	# Video Plays	-0.0093*	0.051
Education	# Chapters	-0.0313	< 1e-15
	# Active Days	-0.0296	< 1e-15
	# Events	-0.0287	< 1e-15
	# Video Plays	-0.0329	< 1.9e-12

TABLE 4.2: The numeric values of the correlation of quasi-identifier frequencies with various activity metrics. Correlation values that are marked with an asterisk are not significant at the $\alpha = 0.05$ level.

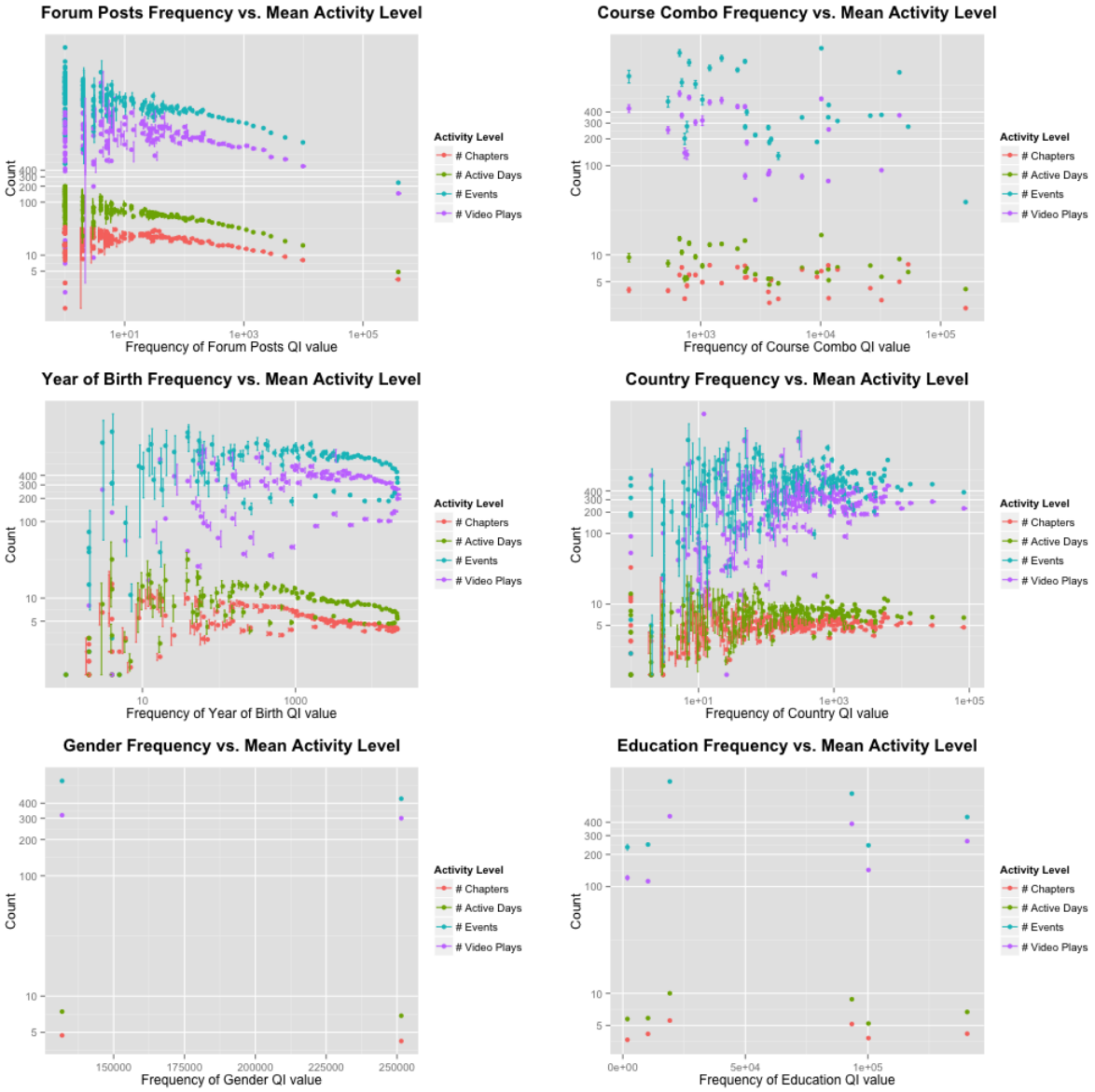


FIGURE 4.3: These graphs show the relationship between the frequency of each of the six quasi-identifier attributes with various activity metrics. As with the other numeric attributes, the strongest relationship again occurs between the frequency of forum posts with each of these values.

4.1.2 Quasi-identifier combination frequencies

Above, we observed a relationship between most quasi-identifier variables with grade, performance, and activity variables, which suggested that the k-anonymization process may ultimately bias the dataset towards lower performers due to the fact that rarer values of these quasi-identifiers tend to be associated with higher values. Here, we analyze whether the rarity of the *combination* of all six quasi-identifiers also has this same negative correlation with grade, performance, and activity variables.

Quasi-identifier combination frequencies versus grade

Below we observe a fairly strong relationship between the frequency of the combination of six quasi-identifiers with the mean grade for each value. The correlation between quasi-identifier frequency and mean grade is significant and has a value of -0.3419, signaling that more frequent quasi-identifier values tend to be associated with lower grades. In conjunction with our earlier finding that each of the individual quasi-identifier values was also individually negatively correlated with grade, this further suggests that k-anonymization will create a negative bias in the reported grades.

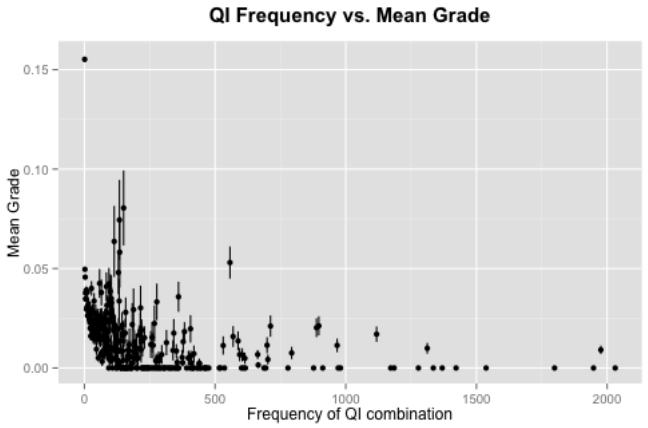
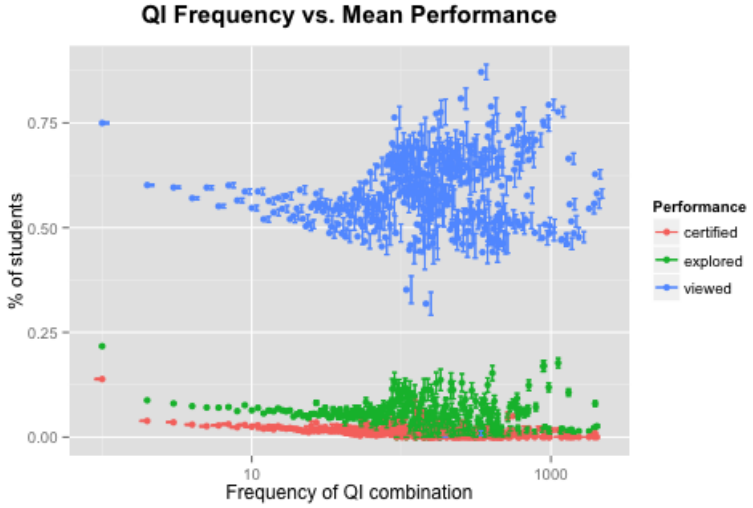


FIGURE 4.4: The relationship between the frequency of quasi-identifier combinations with the mean grade. The correlation of this relationship is -0.3419 with a p-value of 1.043e-09.

Quasi-identifier combination frequencies versus performance

Although we observe no significant relationship between the frequency of quasi-identifier combinations with the percentage of students who view the "courseware" tab, there does exist a negative relationship between the frequency of quasi-identifier values with the percentage of students who access at least half the chapters ("explored"), as well as with who ultimately get certified. As before, the negative sign of these correlations between quasi-identifier frequency with the "explored" and "certified" values suggest that the k-anonymized dataset is likely to have lower performance metrics than the original dataset.



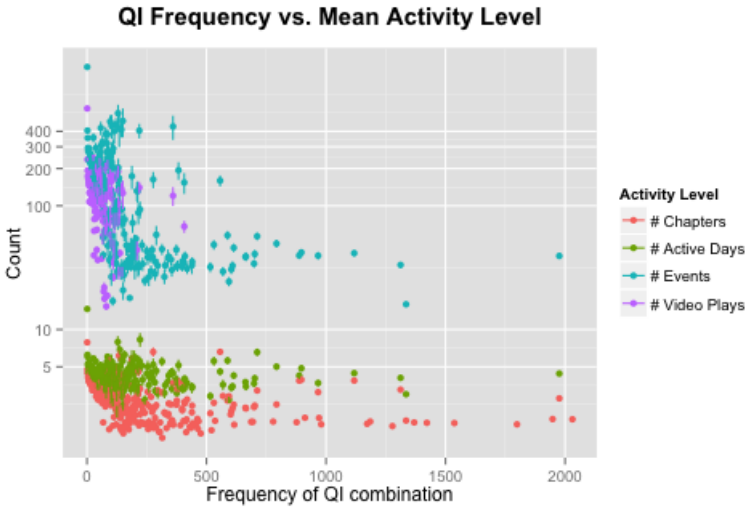
Correlations(QI freq, performance)		
Perf	Correlation	p-value
Viewed	0.0375*	0.5159
Explored	-0.1129	0.0500
Certified	-0.3121	3.027e-08

FIGURE 4.5: Relationship between the frequency of the combination of all quasi-identifier values with performance variables. Note that the relationship with the "viewed" variable is not statistically significant.

Quasi-identifier combination frequencies versus activity

We similarly find a negative relationship between the frequency of quasi-identifier combinations with measures of course activity. Although the relationship between quasi-identifier frequency with the number of active days is insignificant at the $\alpha = 0.05$ level, the relationship between the other activity levels are all fairly strongly negative. As was also seen between the individual quasi-identifier variables' correlation with measures of

activity, this suggests that the de-identification process is likely to skew these variables downward.



Correlations(QI freq, activity)		
Activity	Correlation	p-value
# Chapters	-0.3358	2.272e-09
# Active Days	-0.1239*	0.05951
# Events	-0.3651	1.008e-08
# Video Plays	-0.2585	0.005289

FIGURE 4.6: Relationship between the frequency of the combination of all quasi-identifier values with activity variables. Note that the correlation with the number of active days is insignificant at the $\alpha = 0.05$ level.

4.2 Measures of data utility

In order to test the hypothesis that strong correlations between the frequency of quasi-identifier attributes with the values of non-identifier fields creates a skew in the data, there must be some pre-determined notions of skewness or, more generally, the "utility" of a dataset. Literature that was reviewed in Chapter 2 often defined the "utility" of a dataset in terms of the error of a specific statistical procedure, such as linear regression or classification, and attempted to minimize this error metric during the de-identification process. However, as discussed previously, using this notion of utility can often be problematic in situations where the desired statistical analysis is not known in advance.

For this reason, we opt to use general measures of utility that either measure the change in spread of the data, or that measure the degree to which a given dataset has been "anonymized". Specifically, four measures of utility are studied:

Definition 4.1 (Discernibility metric)**.** This metric penalizes every record proportionally to the number of records that are indistinguishable from it. Intuitively, this captures the desire to be able to distinguish between records – the more indistinguishable that records are from each other, the higher degree that the de-identification "blurred" records together. The discernibility metric is formally defined on a pre-suppression de-identified dataset D as follows:

$$C_{DM}(D,k) = \sum_{E \in |E| \geq k} |E|^2 + \sum_{E \in |E| < k} |D||E|$$

where $|D|$ is the size of the original dataset and the sets E are the equivalence class induced by the k-anonymization [17].

The first term in this metric penalizes the equivalence classes that are present in the de-identified dataset by the size of their equivalence classes, while the second term penalizes the suppressed records by the size of the original dataset, due to the fact that their suppression makes them "indistinguishable" from every record in the dataset [17].

Note that this calculation must be performed before records are suppressed, or else there would never be any records belonging to the $|E| \geq k$ set. After the calculation of this metric, the necessary rows can then be suppressed.

Definition 4.2 (Average equivalence class size)**.** This is the average size of the equivalence classes that are present in the de-identified dataset. If E denotes the set of equivalence classes in the de-identified dataset, $E_1, ..., E_n$, then the average equivalence class is simply calculated as:

$$\frac{\sum_i |E_i|}{|E|}$$

Average equivalence class size is very closely related to the above-defined discernibility metric, except that it does not penalize for suppressed rows [18].

Definition 4.3 (Entropy)**.** Entropy is a measure of how much information is missing from a dataset. Mathematically, it is defined as:

$$H(E) = -\sum_i \frac{|E_i|}{N} \log_2(\frac{|E_i|}{N})$$

where $|E_i|$ is the size of the i^{th} equivalence class, and where N is the total number of rows in the dataset. $\frac{|E_i|}{N}$ can be thought of as the probability of a certain record's membership to a given equivalence class.

Intuitively, entropy can be thought of as the degree to which the information in a dataset identifies a given record. If each record is easily distinguished from every other record, then the $\frac{|E_i|}{N}$ values will be low and the entropy will be low. However, if many records are indistinguishable from each other, then entropy will be high. This can be a very valuable metric in terms of measuring data quality [10].

Definition 4.4 (Q-diversity)**.** Given a dataset D and its de-identified version A, Q-diversity is a metric of the "distance" between its discretized values and its original values, formally defined as:

$$q_d(A_a) = \frac{S(A_a)}{|D_a|}$$

where $S(A_a)$ is a function whose value denotes the number of values of quasi-identifier "a" are present in a given dataset [19].

Given this notion of distance between original and discretized values, an overall notion of utility can then be created, such that:

$$\text{Q-Diversity} = avg(q_d(v_1), ..., q_d(v_n))$$

for each QI, 1 through n.

Intuitively, Q-diversity measures the degree to which discretization of values has blurred the dataset. Unlike the previous three measures of utility, however, *lower* values signal a worse quality dataset [19]. Note that Q-diversity ranges between 0 and 1, with 1 signaling that no discretization has occurred. Since the other three notions of utility are encoded such that higher values are worse, we will report the metric of 1 -(Q-diversity) so that comparisons can more easily be made between the four metrics.

Of the four above-outlined notions of utility, average equivalence class size and the discernibility metric are based on the size of equivalence classes, and therefore are measures of how distinguishable records are based on their quasi-identifiers. Average equivalence class size does not take suppressed rows into account, so is a good measure of the effect of generalization on the size of equivalence classes, whereas the discernibility metric measures the effect of both suppression and generalization. Due to the similarity of these metrics, they are likely to be highly correlated. Although entropy also depends on the size of equivalence classes, it moreso measures the *balance* between sizes rather than the sum or mean of the sizes, and therefore measures a different aspect of the dataset than

the other metrics. Finally, Q-diversity measures the degree of generalization of values, so is unlikely to be directly correlated with any of the above metrics.

With these measures of utility, we are now able to quantify the effect that various factors have on a given dataset, in addition to looking at simple numeric qualities like changes in mean and standard deviation of numeric attributes. Using these tools, we proceed to analyze the effect of various factors on dataset quality.

Specifically, given our hypothesis that the unequal distribution of quasi-identifiers is one of the driving factors of bias and loss of utility in datasets throughout the de-identification process, then we should expect a change in dataset utility when:

- k **is increased.** In this case, the quasi-identifier frequency threshold for dropping a given record becomes higher and therefore, if there exists a correlation between quasi-identifier combination frequency and the mean values of performance variables, then an increase in k should correspond with a loss of utility and an increase in bias.
- **Quasi-identifier variables are eliminated.** If a single quasi-identifier's frequency is highly correlated with a given numeric attribute, then simply dropping the entire quasi-identifier variable should lower the correlation between *combinations* of quasi-identifiers and therefore lessen the bias introduced into the dataset through the de-identification process.
- **Correlation between quasi-identifier rarity with numeric attributes is altered.** In order to support our hypothesis that a high correlation between quasi-identifier frequency and numeric attributes causes skewing of de-identified datasets, we expect that, if the correlation of a certain quasi-identifier variable's frequency were to be increased with a given numeric attribute, the skewness of that attribute would likewise increase (and the utility would decrease).

In the following sections, we analyze the outcomes of these three experiments in order to test the hypothesis that the unequal distribution of quasi-identifier frequencies may introduce skewness into the dataset.

4.3 The effect of k on statistical bias and utility

The choice of k in a k-anonymization framework is a large determinant of the degree of utility loss and gain in bias that is introduced by the anonymization process. Recall

that, in a k-anonymous dataset, any single record must be indistinguishable from at least k-1 distinct individuals in terms of the quasi-identifier fields. For example, in a dataset with "State of Residence" as the single quasi-identifier field, any record belonging to a state with less than k residents must have its value either removed from the dataset or "generalized" to a broader category that contained more than k residents, such as "Midwestern US". Accordingly, greater values of k provide stricter anonymization requirements, because each record is required to be indistinguishable from a greater number of other records, meaning that more suppression and generalization will be required.

Furthermore, legal restrictions on de-identification, and specifically the value of k (or equivalent values in other anonymization schemes), are often open to interpretation. Significant debate among legal specialists was necessary when interpreting the level of anonymization that FERPA laws required of the edX dataset. Even though the final determination was to use a value of $k = 5$, there was considerable discussion whether k may have only been required to be 3.

To investigate the extent of the effect of the choice of k on the datasets, we first ran the k-anonymization procedure on the original dataset for seven different values of k. Summary statistics on the changes induced by different values of k are shown below – it can be seen that larger values of k correspond to smaller de-identified datasets. Furthermore, gender and activity metrics become increasingly skewed from the original dataset's values as k increases.

	Original	k=3	k=4	**k=5**	k=6	k=7	k=8
Number of records	419174	375358	363154	353066	344066	336032	329088
% Male	60.0%	61.0%	61.2%	61.4%	61.5%	61.7%	61.8%
% Viewed	60.0%	58.1%	57.8%	57.5%	57.5%	57.4%	57.4%
% Explored	8.6%	6.7%	6.5%	6.2%	6.2%	6.2%	6.1%
% Certified	3.8%	2.4%	2.2%	2.0%	2.0%	2.0%	1.9%

TABLE 4.3: Table of summary statistics describing changes in the data as the original dataset is de-identified with different values of k. As k increases, the resulting de-identified dataset becomes smaller and more skewed.

Given a k-anonymity framework for de-identification, the existence of a correlation between the frequency of quasi-identifier values with numeric attributes should intuitively drive a relationship between k and loss of data utility. As the value of k increases, grades associated with increasingly "less rare" grades become subject to suppression and generalization. In the presence of a correlation between the rarity of a grade with numeric

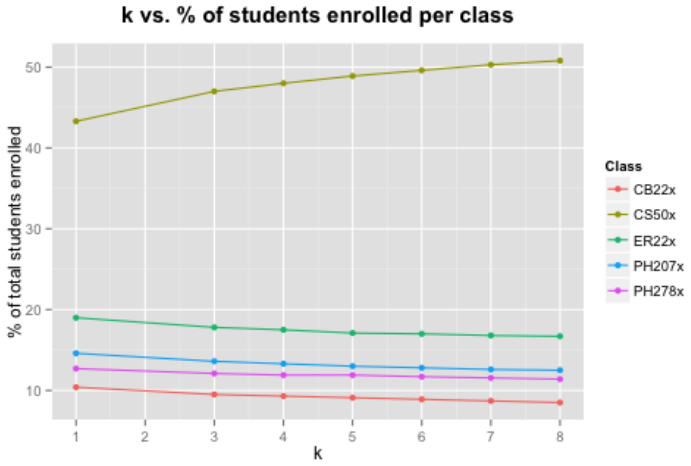


FIGURE 4.7: Change in course enrollments for different values of k in k-anonymization.

attributes, this suggests that numeric attributes may become more skewed upward or downward, depending on the sign of the correlation.

Not surprisingly, we observe that as k increases (and thus as the anonymity requirement becomes stricter, requiring each record to be less and less unique), the mean grade, the mean performance, and the mean activity level all decrease, as was predicted in the previous section by the analysis of the rarity of certain quasi-identifier values versus the grade, performance, and activity metrics.

Not only are the means of certain numerical attributes skewed downward by an increasing value of k, but so is the utility of the entire dataset, in terms of each of the four utility measures that were defined above, as is seen in the below table.

k	**DM**	**Avg(| EqCl |)**	**Entropy**	**1-Qdiv**
k=1	7.83×10^7	4.27	51.6	0
k=3	1.85×10^{10}	19.0	159	0.473
k=4	2.36×10^{10}	23.1	183	0.473
k=5	2.78×10^{10}	26.8	202	0.479
k=6	3.16×10^{10}	30.2	219	0.480
k=7	3.5×10^{10}	32.7	230	0.474
k=8	3.79×10^{10}	35.8	243	0.479

TABLE 4.4: As k increases, there is a loss of data utility in terms of every utility metric, signaling a decrease in data utility as the data must comply to stricter de-identification requirements.

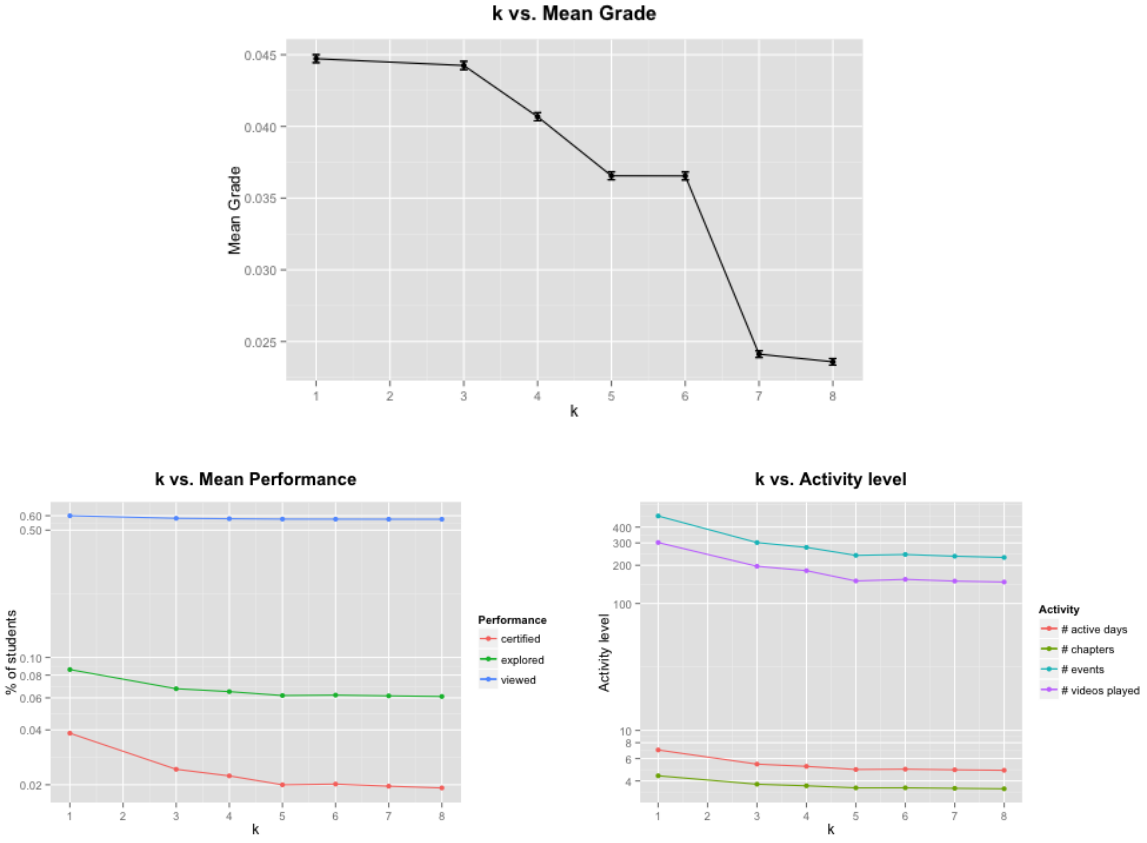


FIGURE 4.8: As k increases in a k-anonymity framework, we observe a decrease in the mean grade, performance, and activity levels, likely due to the association between rarer quasi-identifier values with higher values of these variables.

As k increases, we therefore have seen that bias is introduced into the dataset and utility is decreased. This observation is in accordance with our hypothesis that the correlation of quasi-identifier frequencies with numeric attributes may contribute to the bias and decreased utility of datasets. To further explore this hypothesis, we will also explore the relationship between the correlation of quasi-identifier characteristics and other effects it may have on the resulting de-identified dataset.

4.4 The effect of suppressing entire quasi-identifier columns on statistical bias

Under the hypothesis that a high correlation between the frequency of quasi-identifier attributes with numeric attributes is the cause of skewness during the de-identification process, we would expect that the complete omission of a quasi-identifier column whose values are highly correlated with a numeric attribute would reduce the bias of that attribute. Below, we perform k-anonymization on the original dataset, using only 5 of

the 6 QI fields at once, in order to observe the effect of omitting a given QI field on the skewness of certain numeric attributes in a dataset. Intuitively, we expect that, the higher the absolute value of correlation between the rarity of the values of the removed column with a given attribute, the less skewed the overall dataset will be in terms of that numeric attribute.

The tables below compare the correlation between the QI frequency and grade, performance, and activity metrics, as compared with the resulting mean of these columns when a given QI attribute is omitted. We find that, when the number of forum posts (which has the highest absolute value of correlation with each of the numeric attributes) is deleted as a quasi-identifier attribute, the resulting mean of every column except the number of chapters is the highest. This fits with our hypothesis: since records with high values are no longer being suppressed due to their unique number of forum posts, the mean of associated performance and activity metrics becomes higher in the de-identified dataset. However, the relationship between the removal of the other quasi-identifier variables with the mean grade of the dataset does not appear to relate to the correlation of the frequency of the QI with grade.

QI removed	Cor(QI freq,grade)	Grade
None-Original	NA	0.045
None-Deidentified	NA	0.037
Forum posts	-0.43	0.061
Course combo	-0.17	0.024
Country	-0.051	0.037
Year of birth	-0.045	0.043
Gender	-0.045	0.029
Education	-0.023	0.039

QI removed	Cor(QI freq, viewed)	Viewed	Cor(QI freq, explored)	Explored	Cor(QI freq, certified)	Certified
None-Original	NA	0.60	NA	0.086	NA	0.038
None-Deidentified	NA	0.58	NA	0.062	NA	0.020
Forum posts	-0.21	0.60	-0.35	0.084	-0.39	0.037
Course combo	0.048	0.43	-0.085	0.066	-0.14	0.018
Country	-0.050	0.59	-0.047	0.076	-0.042	0.031
Year of birth	-0.0017	0.59	-0.048	0.080	-0.044	0.034
Gender	0.030	0.58	0.00090	0.066	-0.033	0.024
Education	-0.032	0.58	-0.029	0.068	-0.021	0.025

QI removed	Cor(QI freq, days)	# Active days	Cor(QI freq, chapters)	# Chapters
None-Original	NA	7.0	NA	4.4
None-Deidentified	NA	4.9	NA	3.5
Forum posts	-0.47	6.8	-0.43	4.3
Course combo	-0.12	5.1	-0.2601	4.8
Country	-0.032	6.1	-0.046	4.1
Year of birth	-0.10	6.5	-0.083	4.2
Gender	-0.025	5.3	-0.046	3.7
Education	-0.030	5.5	-0.031	3.8

QI removed	Cor(QI freq, events)	# Events	Cor(QI freq, vid plays)	# Video Plays
None-Original	NA	489	NA	302
None-Deidentified	NA	240	NA	151
Forum posts	-0.41	470	-0.26	300
Course combo	-0.14	204	-0.0171	86
Country	-0.041	396	-0.040	258
Year of birth	-0.071	436	-0.061	278
Gender	-0.054	290	-0.0093	191
Education	-0.029	305	-0.033	195

TABLE 4.5: This table explores the relationship between the mean grade, mean performance levels, and mean activity levels between QI rarity in datasets in which one QI variable has been omitted. In every variable except for the number of chapters, we see that when the number of forum posts (which, for every variable, has the highest correlation between its frequency with values of the numeric attributes) is omitted as a QI variable, the resulting mean of the numeric attribute is highest. This supports the hypothesis that a high negative correlation between QI frequency with a given numeric attribute may skew the variable downward during the de-identification process. Although we do observe that the variable with the highest correlation between QI rarity and grade also tends to have the highest mean of numeric attributes, there does not appear to be a strong relationship present between the other variables.

From these analyses, it therefore appears that there may exist a benefit in completely omitting quasi-identifier attributes from a dataset if the rarity of their values is highly correlated with many numeric attributes. In doing so, the loss of information caused by the exclusion of the attributes must be taken into account. The number of forum posts, for example, may not be too important of a quasi-identifier attribute to be kept in the dataset, since it does not provide valuable demographic information and is a measure of activity that is correlated with other *non*-quasi-identifier measures of activity (like the number of video plays, for example). On the other hand, if "gender" or "year of birth" were to be excluded, this may significantly hinder analyses of student performance based on demographic traits, which may be more undesirable than a slightly skewed dataset.

The suppression of quasi-identifier columns whose *rarity* of values is highly correlated with numeric attributes has the same effect as decreasing the amount of sample bias. Thus, an improvement in sampling methods in order to ensure the balance of background covariates (and thereby quasi-identifier attributes) is another way to reduce the amount of bias introduced to a dataset.

4.5 The effect of the correlation of QI rarity with grade on statistical bias

So far, we have observed that 1. as k is increased, the amount of bias that is introduced by the k-anonymization process increases, and 2. the suppression of the forum posts column, whose frequency has the highest negative correlation with performance and activity metrics, resulted in the highest increase in the de-identified mean of performance and activity metrics. Both of these observations support our hypothesis that, as the absolute value of the correlation between the frequency of a given quasi-identifier field's occurrence increases with a non-sensitive column (such as grade), the skew of that non-sensitive column will increase.

In order to concretely test this hypothesis further, we now alter given quasi-identifier fields in order to have different correlations with numeric attributes and then explore the skewness of the resulting de-identified datasets. Due to the heavy skew of grades towards zero in this dataset, we expect that a highly negative correlation between forum post value frequency will create a higher degree of negative bias in grade than a highly positive correlation will create a positive bias.

We perform this analysis on a random subset of about a quarter of the total rows in the original dataset – 100,000 rows. Then, a vector is generated whose values represent the frequencies of occurrence of a randomly-generated number of forum posts whose sum is constrained to equal the number of rows in the dataset. Then, each of these frequencies are assigned to a given record in multiple permutations in order to generate different simulated correlations between the frequency of a given value of the number of forum posts with grade. These datasets are then k-anonymized with k=5, and the bias of the grade variable for each of these different is then analyzed.

The first graph below reports the relationship between the entropy of the grade vector after k-anonymizaton (where $k = 5$) and the different correlated frequencies. Here we see that when correlation between the forum post frequency and grade is low, then the entropy of the data is high, suggesting that high grades are being dropped, which creates a more homogeneous population of grades. As the correlation between the forum post frequency and grade increases, the entropy decreases, which suggests that fewer unique values are being dropped.

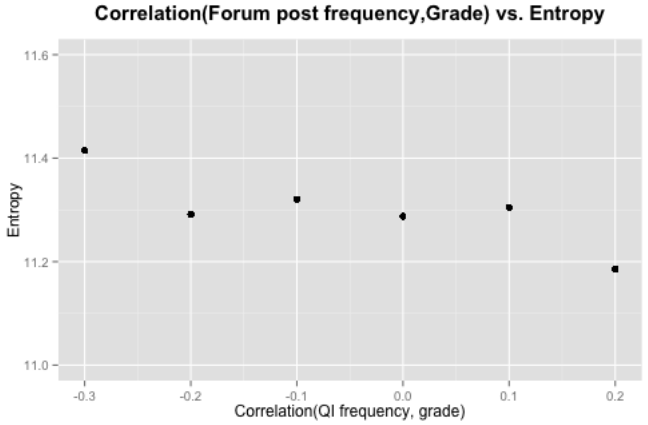


FIGURE 4.9: This graph shows the relationship between the correlation of a specific quasi-identifier attribute with grade versus the entropy of the entire dataset. We see, in general, that the entropy decreases as the correlation becomes higher. This can be explained by he fact that increasingly positive correlations correspond to less rare values becoming dropped, and therefore a lower degree of information loss introduced by the de-identification process.

We also observe that the mean of the anonymized dataset is skewed most severely downward when the correlation is the most negative, and then approaches the true mean of the dataset as the correlation increases. This is because a highly negative correlation means that rarer QI values are more associated with higher values of grade, meaning that these high grades are more likely to be dropped from the dataset and therefore decrease the overall mean of the dataset.

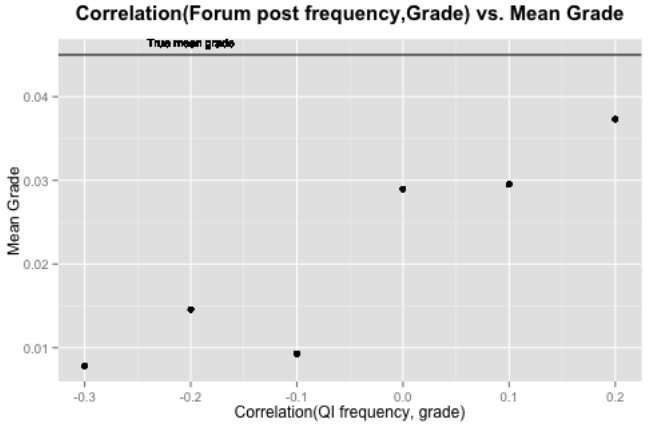


FIGURE 4.10: As expected, a relationship is seen in which, the more negative the correlation between quasi-identifier frequency with grade, the more negative the bias in terms of the mean grade in the de-identified dataset. This is explained by the fact that a negative correlation corresponds to high values of grade being associated with rare values of the number of forum posts, meaning that these records with high grades are more likely to be dropped from the dataset.

4.6 The effect of generalization versus suppression on statistical bias

We have now observed through three different analyses that the unequal distribution of quasi-identifier frequencies may introduce bias into a dataset during the de-identification process. However, one factor whose effect on bias we have not yet explored is the balance between generalization and suppression in de-identifying the dataset. In order to faithfully replicate the anonymization process used to de-identify edX data for public release in 2014, all above analyses employed a "suppression-emphasis" approach toward k-anonymization. In this approach, the names of the countries were first generalized to region or continent names, then date-time stamps were transformed into date stamps, and finally any existing rows that were not k-anonymous after these generalizations became suppressed. In the process, records were removed whose date of birth corresponded to years before 1931.

However, it is reasonable to question the effect that adjustments to the balance between generalization and suppression may have on the bias introduced into certain numeric attributes. Generalization of attributes decreases the need for the suppression of records by monotonically increasing the size of each equivalence class and thus decreasing the chance that a record will need to be suppressed in order to satisfy k-anonymity. However, the inferences that can be drawn from generalized values are often less powerful than those that can be drawn from more granular values – for example, correlations may be hard to calculate with binned numeric attributes.

Since our above analysis indicated that the uneven distribution of the number of forum posts may be a contributing factor to the introduction of bias into the de-identified edX dataset, we explore the effect of generalization of this attribute, where k in k-anonymity is maintained at 5. Most basically, we observe that as the bin size for the number of forum posts increases, there is an increased range of values of the number of forum posts that are represented in the resulting de-identified dataset. This can be seen in the below table, where the upper bound for the number of forum posts increases from 7 to 17 as the bin size increases from 1 to 6. This can be explained by the fact that broader bins result in a higher number of records with given values of the forum post quasi-identifier, and this decreases the chance of associated records from being suppressed from the dataset.

Bin size	**Unique forum post values**
1	0, 1, 2, 3, 4, 5, 6, 7
2	0-1, 2-3, 4-5, 6-7, 8-9
3	0-2, 3-5, 6-8, 12-14
4	0-3, 4-7, 8-11, 12-15
5	0-4, 5-9, 10-14
6	0-5, 6-11, 12-17

TABLE 4.6: As the bin size for the number of forum posts increases, there is an increased range of values of the number of forum posts that is represented in the resulting dataset.

Below is a table reporting summary statistics for de-identified datasets that result after binning forum posts with bin sizes ranging from 1 (i.e., the original values) through 6 (i.e., 0-5, 6-11, etc.).

	# Forum Posts: Bin Size					
	1	2	3	4	5	6
Number of records	352278	360420	364499	366736	368401	369618
% Male	61.4%	61.2%	61.1%	61.0%	60.9%	60.9%
% Viewed	57.5%	58.3%	58.7%	58.9%	59.1%	59.2%
% Explored	6.2%	6.6%	6.9%	7.1%	7.2%	7.4%
% Certified	2.0%	2.3%	2.5%	2.7%	2.8%	2.9%

TABLE 4.7: Table of summary statistics describing changes in the data as the original dataset is de-identified with different bin sizes for the number of forum posts, all at the $k = 5$ anonymity level.

We observe that the dataset size increases as the bin size increases, which is consistent with the observation that generalization increases the size of equivalence classes and therefore decreases the number of rows that must be suppressed. Furthermore, we notice that performance metrics also increase as the bin size increases, suggesting that generalization may alleviate bias by preventing records associated with rarer quasi-identifier values from becoming suppressed.

In order to confirm this hypothesis, we plot the relationship between the bin size of the number of forum posts with mean grade, performance, and activity levels, as shown below.

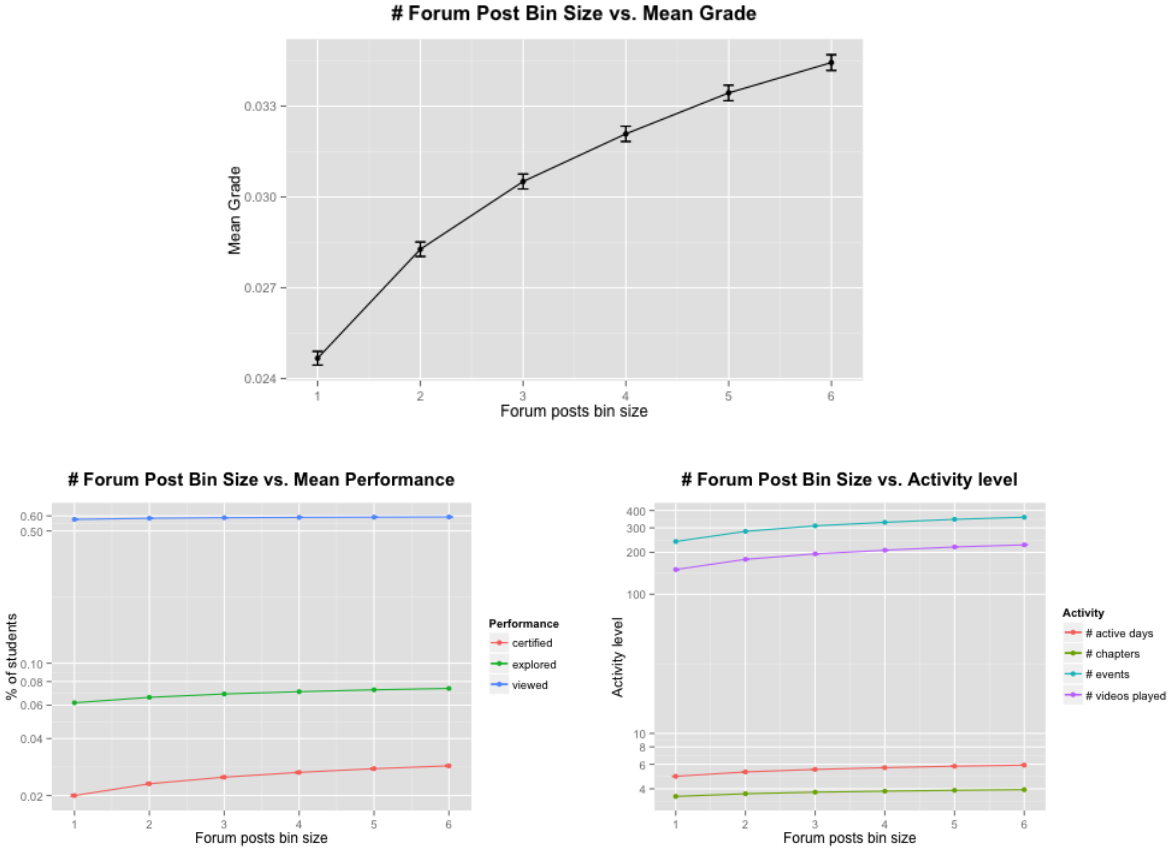


FIGURE 4.11: As the size of the forum post bin size increases in a k-anonymity framework, we observe an increase in the mean grade, performance, and activity levels, likely due to the fact that the generalization allows more values to be included in the de-identified versions. Thus, this may suggest that above findings regarding the bias introduced by de-identification can be counteracted by using a more generalization-heavy anonymization approach.

As expected, as the amount of generalization increases (i.e., as forum post bin size increases), we see that the mean grade, performance, and activity levels all become less biased, and thus approach their true values. This can be explained by the fact that the binning increases the representation of each equivalence class, in effect increasing the threshold at which records associated with higher grades and performance were cut off.

If increased generalization is associated with decreased bias, then why would a k-anonymization approach that emphasizes suppression ever be employed? In order to understand the tradeoffs that are involved in using generalization over suppression, we measure the utility of de-identified datasets that are produced using multiple bin sizes. The results are shown in the below table.

Bin size	**DM**	**Avg(| EqCl |)**	**Entropy**	**1-Qdiv**
1	2.81 x 10^{10}	27.0	203	0.479
2	2.47 x 10^{10}	27.4	206	0.481
3	2.30 x 10^{10}	27.5	208	0.483
4	2.21 x 10^{10}	27.7	209	0.481
5	2.14 x 10^{10}	27.8	210	0.482
6	2.08 x 10^{10}	27.8	210	0.482

TABLE 4.8: As the bin size for the number of forum posts increases, there is a loss of data utility in terms of average equivalence class size and entropy, but an increase in the data utility in terms of the discernibility metric.

Recalling that, for each of the four reported utility metrics, higher values are associated with lower utility, we observe that increased bin sizes are associated with an improved discernibility metric but decreased utility in terms of average equivalence class size and entropy. This follows from the fact that the discernibility metric heavily penalizes for suppressed rows, so increasing the amount of generalization necessarily decreases the amount of suppression and thus improves the ability to distinguish between different records.

However, the average equivalence class size is worsened as generalization increases because, among non-suppressed records, the number of records with a given set of quasi-identifiers becomes larger and therefore it is harder to associate a given record with precise quasi-identifier values, which decreases the utility of the dataset. Similarly, since equivalence class sizes are increased, there is a decreased ability to characterize a record from other records, and therefore the dataset's entropy decreases. Therefore, we observe that the utility of the dataset is decreased in the sense that it becomes more difficult to distinguish between records in terms of their quasi-identifier characteristics.

Stemming from these findings, we also note that generalization makes it difficult to draw statistical conclusions from a dataset due to its discretization of numeric attributes. The mean of a column that has undergone generalization can be maintained by computing a weighted mean of the pre-discretized values and then reporting this value as the "bin average" in the de-identified dataset. By averaging these bin averages, the resulting mean will represent the true mean of the pre-discretized values.

Such a solution cannot be easily derived for two-dimensional relationships between generalized values, however. Consider the below table, which reports the correlations between the number of forum posts in the original dataset with various numeric attributes in the left column, and then compares this to the correlation between the "bin average" value for various bin sizes with the same numeric attributes.

Correlations of forum posts with numeric attributes							
		Bin size					
	Original	1	2	3	4	5	6
Grade	**0.159**	0.105	0.0980	0.0919	0.0833	0.0732	0.0533
Viewed	**0.0444**	0.0683	0.0582	0.0462	0.0372	0.0294	0.0228
Explored	**0.127**	0.0744	0.0710	0.0661	0.0620	0.0554	0.0418
Certified	**0.152**	0.0868	0.0810	0.0758	0.0699	0.0598	0.0482
# Active Days	**0.236**	0.117	0.111	0.106	0.0940	0.0855	0.0649
# Chapters	**0.154**	0.143	0.127	0.115	0.100	0.0858	0.0715
# Events	**0.283**	0.103	0.103	0.0964	0.0986	0.0913	0.0597
# VIdeo Plays	**0.0929**	0.0943	0.105	0.103	0.125	0.110	0.0683

TABLE 4.9: This table describes the change in correlations between the number of forum posts and other numeric attributes as the degree of generalization of the number of forum posts is increased. Note that the first column represents the true correlations in the original dataset. All of these analyses hold are for k-anonymization where $k = 5$.

Strikingly, the correlation between the number of forum posts with every numeric attribute except for the `viewed` attribute becomes more biased as the number of bins increase. Noting that the column corresponding to a bin size of 1 represents the "suppression emphasis" k-anonymization approach, and that its correlations are consistently the closest to the original correlations, this suggests that generalization distorts aspects of joint relationships between variables through its discretization of values.

Thus, we have encountered the fundamental tradeoff between generalization and suppression that was discussed earlier – although an approach emphasizing suppression may introduce bias in a given numeric attribute if there exists a quasi-identifier whose frequency is correlated with the numeric attribute, generalization may also inherently distort datasets in terms of correlational and other multidimensional relationships.

One potential improvement to generalization may be to more evenly distribute the bin sizes, using fine bucket sizes for values that are well-represented and using larger bucket sizes for less well-represented values. We create a dataset with mixed bucket sizes, generalizing the number of forum posts into bins of size five for values above 10, to be consistent with the schema that was used in the "generalization emphasis" anonymization approach employed by the edX team. The resulting de-identified dataset under this mixed generalization schema had the following unique values of the number of forum

posts: 0, 1, 2, 3, 4, 5, 6, 7, 11-15. Note that representation for 8, 9, and 10 forum posts was low enough that all such records had to be suppressed. We also analyzed the utility of the resulting dataset with mixed bin sizes, and report the values below.

Bin size	**DM**	**Avg(| EqCl |)**	**Entropy**	**1-Qdiv**
1	2.815 x 10^{10}	27.0	203	0.479
Mixed	2.814 x 10^{10}	27.0	203	0.478

TABLE 4.10: A non-uniform bin size (i.e., bin size of 1 for values below 11 and bin size 5 above 11) is characterized by very similar utility values than the de-identified dataset that results from simply using a bin size of 1.

In terms of utility, we see that this resulting dataset is quite comparable to the de-identified dataset that only used a bin size of 1. The mixed bin sizes therefore do not provide a significant improvement in terms of the number of suppressed rows (as seen by the relatively high discernibility metric), but they are advantageous in the fact that they do not cause an increase in the average equivalence class size or in entropy, meaning that records remain fairly distinguishable from each other in terms of their quasi-identifier characteristics.

Correlations of forum posts with numeric attributes			
		Bin size	
	Original	1	Mixed
Grade	**0.159**	0.105	0.990
Viewed	**0.0444**	0.0683	0.0635
Explored	**0.127**	0.0744	0.0700
Certified	**0.152**	0.0868	0.0815
# Active Days	**0.236**	0.117	0.110
# Chapters	**0.154**	0.143	0.138
# Events	**0.283**	0.103	0.0950
# VIdeo Plays	**0.0929**	0.0943	0.0943

TABLE 4.11: This table describes the change in correlations between the number of forum posts and other numeric attributes with mixed bin sizes as compared with no generalization (i.e., bin size of 1). Note that the first column represents the true correlations in the original dataset. All of these analyses hold are for k-anonymization where $k = 5$.

Similarly, this table reveals that the de-identified dataset that uses mixed bin sizes has correlations that are farther from the true values than the de-identified dataset that does not bin the number of forum posts (i.e., where bin size is 1), but that are significantly closer to the true correlation values than datasets that use consistently larger bin sizes (i.e., bins of size 2 or more).

This suggests that using bin sizes that optimize for equal numbers of records within each bin may provide a compromise between the loss of utility and the distortions caused in numeric analysis like correlations between different variables.

Chapter 5

Conclusions and future directions

This paper has primarily explored the relationship between quasi-identifier characteristics as a possible source of bias and loss of utility during the k-anonymization process. By requiring that each equivalence class have a minimum size of at least k, k-anonymization decreases the ability for adversaries to re-identify records by joining them with outside datasets. In doing so, however, there is a chance that bias is introduced into the dataset by unevenly suppressing records with certain characteristics.

We hypothesized that, in particular, the strength of the correlation between numeric attributes and the frequency of occurrence of quasi-identifier values was a large determinant of how biased a given de-identified dataset would become. Visual explorations of de-identified data available through a massive online open course platform, edX, suggested that a relationship between quasi-identifier frequency did indeed exist with other numeric attributes.

By modifying factors that affect the relationship between correlations of quasi-identifier value frequencies with other numeric attributes, we confirm the possibility that correlations between the frequency of quasi-identifier attributes with other numeric attributes are a contributing source of bias and loss of utility. Specifically, we investigated three factors in their contributions to this bias.

I. **Increasing the value of k in k-anonymity.** In a situation where a quasi-identifier field's frequency of values is correlated to a numeric attribute, increasing the value of k is akin to changing the quasi-identifier rarity threshold at which a given record must be cut. If a quasi-identifier's rarest values are more often tied to either high or low numeric attributes, this therefore would suggest that higher

levels of k (which correspond to greater anonymity requirements) would create biases in the resulting de-identified dataset's numeric attributes. We did indeed witness this trend: higher values of k corresponded to downward skews in terms of numeric attributes, caused by the negative correlation between quasi-identifier fields' frequency of occurrence with numeric attributes. This suggests that the combination of correlated quasi-identifier field frequencies with high levels of k may be particularly dangerous in introducing bias into a dataset.

II. **Eliminating quasi-identifier fields with high correlations of value frequency to numeric attributes.** Given a quasi-identifier field with a high correlation between the frequency of its values with numeric attributes, there may be value in simply omitting the entire field rather than allowing it to create bias within the dataset. In this case, a balance must be maintained between the value of the information encoded in the quasi-identifier field versus the bias created in the dataset.

   It appears from initial findings that a weak relationship between the correlation of a given quasi-identifier field's value frequency with a numeric attribute may exist. Specifically, omitting quasi-identifier fields whose rarity correlation with a numeric field has an absolute value of above 0.25 appeared to correspond the most with noticeable improvements in the bias introduced by de-identification.

III. **Increasing the correlation between quasi-identifier value frequency with given numeric attributes.** The manual alteration of quasi-identifier values confirmed our hypothesis that the amount of bias introduced during the de-identification process may be related to the magnitude of the correlation between quasi-identifier value frequency with that attribute. Due to the fact that numeric attributes like grade in our dataset are highly skewed toward values near 0, this meant that situations in which rare quasi-identifier values are associated with high values of grade caused the most bias in the data. (Naturally, however, if the grades had been skewed toward higher values, rare quasi-identifier values' association with *low* values of grade would have caused the most bias in the data.)

All three of the experiments performed above support the hypothesis that high correlations between quasi-identifier value frequency with numeric attributes is a cause of the introduction of bias within those fields in datasets. Furthermore, the results lend interesting insights into what factors may be responsible for the introduction of bias into a dataset. For example, suppressing the number of forum posts quasi-identifier column had a much greater impact on the mean grade of the anonymized dataset (from 0.045 to 0.061, a change of +0.016) than did increasing the value of k in k-anonymization from 1 to 5 (which changed the mean grade from 0.045 to 0.036, a change of -0.009). On

the other hand, suppression of other columns did not have as great an impact on the mean of numeric attributes as did the increase in k (which corresponds to an increase in the suppression of rows). The interaction between these factors would be a powerful relationship to quantify. Similar analyses could also be done with regards to the effect of adding noise to rows versus adding noise to columns.

Importantly, all of the experiments used a k-anonymization scheme with a given balance of suppression versus generalization, in which countries were generalized into continent or region names, time stamps were generalized to dates, and all remaining records that did not satisfy k-anonymity were suppressed. This approach that emphasizes suppression may have caused a disproportionate bias in resulting activity and performance metrics than would a generalization-emphasizing approach.

To this end, our analyses showed that increased generalization decreased the bias introduced into grade, activity, and performance metrics. However, this increase in generalization was also associated with a loss in data utility as measured by the average equivalence class size and by entropy, due to the fact that generalization blurs the association between records and their quasi-identifier characteristics. Furthermore, the larger the bin size of a generalized attribute, the less accurate statistical analyses with that attribute become, such as its correlation with other columns. This can also affect the ability of researchers to perform analyses like linear regression using generalized attributes. Further research into the effect of anonymization schemes that emphasize suppression versus generalization may be able to shed more light onto how to find an optimal balance between the two.

Another future direction for research that may alleviate the bias of these numeric attributes is the use of *swapping* of quasi-identifier values. Swapping of quasi-identifier rows can solve two problems. First, it prevents the need for rows to be suppressed and therefore lessens the amount of bias that is introduced into the dataset via suppression. Second, it provides a way to reduce the correlation between certain quasi-identifier's value frequencies with numeric attributes, therefore lessening the amount of bias that is introduced when the final k-anonymization step is performed.

Valuable findings may also come from the performance of similar analyses on other datasets. For example, the edX dataset's numeric attributes tended to be highly skewed toward low performers – it may be interesting to perform the same analyses for datasets

whose numeric attributes are more evenly distributed. Furthermore, it would be interesting to observe how different the resulting analyses would be on datasets with fewer or greater numbers of quasi-identifier attributes. With different numbers of quasi-identifiers, the degree of the effect of a high correlation between a *single* quasi-identifier's frequency with another numeric column may be changed.

Of course, all of these analyses have been performed with a focus on minimizing the bias introduced into datasets. However, bias may not always be the ideal error metric to optimize for – for example, a dataset where suppression removes the records with the ten lowest grades might experience an upward bias in the mean grade, but might retain the two-dimensional relationships between grades with other attributes that allow accurate linear regressions to be fit to the data. By this same logic, certain modifications that are made in order to increase the integrity of the de-identification process (by some definition of integrity that a researcher would like to optimize for) may be more amenable to allow researchers to have a more accurate measure of the accuracy of their resulting analyses, through measures like confidence intervals or standard errors.

Clearly, there exist many opportunities for future research that can optimize the k-anonymization process for certain tasks. More generally, these modifications to the de-identification process enable data to be increasingly used in a way that protects the privacy of the individuals whose data is being shared, but that also provides utility to researchers who will be using it.

Bibliography

[1] US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. *http://www.hhs.gov/*, may 2014.

[2] Jon P Daries, Justin Reich, Jim Waldo, Elise M Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9):56–63, 2014.

[3] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. *Mortality*, 1(1.15):1–20, 2014.

[4] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. ACM, 2009.

[5] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems (TODS)*, 33(3):17, 2008.

[6] Benjamin CM Fung, Ke Wang, Lingyu Wang, and Mourad Debbabi. A framework for privacy-preserving cluster analysis. In *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*, pages 46–51. IEEE, 2008.

[7] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[8] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[9] Xiaokui Xiao and Yufei Tao. Dynamic anonymization: accurate statistical analysis with privacy preservation. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 107–120. ACM, 2008.

[10] Jiuyong Li, Jixue Liu, Muzammil Baig, and Raymond Chi-Wing Wong. Information based data anonymization for classification utility. *Data & Knowledge Engineering*, 70(12):1030–1045, 2011.

[11] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 754–759. ACM, 2006.

[12] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380. ACM, 2009.

[13] Keng-Pei Lin and Ming-Syan Chen. On the design and analysis of the privacy-preserving svm classifier. *Knowledge and Data Engineering, IEEE Transactions on*, 23(11):1704–1717, 2011.

[14] Jon Daries. De-identification code. https://github.com/harvard/de_id, 2012. Accessed: 2014-12-22.

[15] Li Xiong and Slawek Goryczka. Data anonymization - generalization algorithms. http://www.mathcs.emory.edu/~lxiong/cs573_s12/share/slides/0131_generalization_slawek.pdf, 2012. Accessed: 2014-01-28.

[16] MITx and HarvardX. Harvardx-mitx person-course academic year 2013 de-identified dataset, version 2.0. *Harvard Dataverse Network*, may 2014.

[17] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[18] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790. ACM, 2006.

[19] Grigorios Loukides and Jianhua Shao. Data utility and privacy protection trade-off in k-anonymisation. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pages 36–45. ACM, 2008.